arXiv:2302.08591v1 [cs.HC] 16 Feb 2023

复杂日常活动、国家层面的多样性与智能手机感知：丹麦、意大利、蒙古、巴拉圭和英国的一项研究

Karim Assi[†]
EPFL
瑞士

Lakmal Meegahapola[†§]
Idiap研究所 & EPFL
瑞士

William Droz
Idiap研究所
瑞士

Peter Kun
奥尔堡大学
丹麦

Amalia de Götzen
奥尔堡大学
丹麦

Miriam Bidoglia
伦敦政治经济学院和政治学, UK

Sally Stares
伦敦大学城市学院
UK

George Gaskell
伦敦政治经济学院
政治学，英国

Altangerel Chagnaa
蒙古国立大学
蒙古

Amarsanaa Ganbold
蒙古国立大学
蒙古

Tsolmon Zundui
蒙古国立大学
蒙古

Carlo Caprini
U-Hopper
意大利

Daniele Miorandi
U-Hopper
意大利

Jose Luis Zarza
亚松森圣母天主教大学 "我们的圣母升天", 巴拉圭

Alethia Hume
亚松森圣母天主教大学，巴拉圭

Luca Cernuzzi
亚松森圣母天主教大学，巴拉圭

Ivano Bison
特伦托大学
意大利

Marcelo Dario Rodas Britez
特伦托大学
意大利

Matteo Busso
特伦托大学
意大利

Ronald Chenu-Abente
特伦托大学
意大利

Fausto Giunchiglia
特伦托大学
意大利

Daniel Gatica-Perez[§]
Idiap研究所 & EPFL
瑞士

摘要

智能手机通过活动识别帮助人们了解人类行为，从而支持人们的日常生活。以往研究侧重于使用惯性传感器检测简单活动（坐姿活动、步行、跑步等），且大多在同质化一个国家内的人群。然而，随着远程/混合办公与学习环境日益普遍，在后疫情时代，人们的久坐行为更多，因此，检测简单活动对于情境感知应用的意义有所降低。因此，目前对于以下问题的理解仍然有限：*(i)* 如何利用多模态智能手机传感器和机器学习模型检测复杂日常活动，从而更好地了解人们的日常生活；以及*(ii)* 模型如何泛化到未见过的国家。我们分析了来自五个国家（意大利、蒙古、英国、丹麦和巴拉圭）637 名大学生的真实环境中的智能手机数据和 ~216K 份自我报告。随后，我们定义了一项包含 12 个类别的复杂日常活动识别任务，并采用不同方法评估其性能。我们发现，尽管通用的多国方法取得了 0.70 的 AUROC，但国家特定方法表现更好，AUROC 分数为 [0.79-0.89]。我们认为，沿着

[†]两位作者贡献相同，按姓氏字母顺序排列。[§]通讯作者为 Lakmal Meegahapola（lakmal.meegahapola@epfl.ch）和 Daniel Gatica-Perez（daniel.gatica-perez@epfl.ch）。

允许为个人或课堂使用而免费制作本作品的全部或部分内容的数字副本或纸质副本，前提是这些副本不得用于营利或商业利益，也不得为此而分发，并且副本须附上本声明及首页的完整引文。对于非作者拥有的本作品组件，必须尊重其版权。注明出处后可摘录。以其他方式复制、再版、发布到服务器或分发至列表，均须事先获得明确许可和/或支付费用。请通过 permissions@acm.org 申请许可。*CHI '23, 2023年4月23日至28日*，德国汉堡© 2023年版权归所有者/作者所有。出版权授权给 ACM。ACM ISBN 978-1-4503-9421-5/23/04...$15.00
https://doi.org/10.1145/3544548.3581190

多样性意识这一方向开展研究，对于通过智能手机和机器学习推进人类行为理解、提升技术在各国真实场景中的实用性至关重要。

CCS 概念

• **以人为中心的计算** → **普适与移动计算实证研究**；**智能手机**；**HCI 实证研究**； • **社会与专业主题** → **地理特征**；**文化特征**。

关键词

被动感知，智能手机感知，情境感知，多样性感知，模型泛化，活动识别，日常生活复杂活动，行为识别，分布偏移，领域偏移

ACM参考文献格式：

Karim Assi、Lakmal Meegahapola、William Droz、Peter Kun、Amalia de Götzen、Miriam Bidoglia、Sally Stares、George Gaskell、Altangerel Chagnaa、Amarsanaa Ganbold、Tsolmon Zundui、Carlo Caprini、Daniele Miorandi、Jose Luis Zarza、Alethia Hume、Luca Cernuzzi、Ivano Bison、Marcelo Dario Rodas Britez、Matteo Busso、Ronald Chenu-Abente、Fausto Giunchiglia 和 Daniel Gatica-Perez。2023年。复杂日常活动、国家层面的多样性与智能手机感知：对丹麦、意大利、蒙古、巴拉圭和英国的研究。载于*2023*年*CHI*人机交互系统大会论文集（*CHI '23*），*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，美国纽约州纽约市，美国，23页。
https://doi.org/10.1145/3544548.3581190

1 引言

近年来，活动识别因其在医疗保健 [113], 体育 [138], 交通运输 [79], 和人类福祉 [70] 等多个领域的实用价值而备受关注。例如，健身追踪类移动健康应用可让用户查看特定活动的指标 [110, 138]。同样，智能家居系统能够根据收集到的人们活动信息，对环境（如照明、温度）进行调整 [64, 80]。情境感知是手机用户体验的一个关键方面，而活动识别的集成使其得以实现 [85, 125]。

传统上，基于传感器的活动识别依赖于安装在人体上的定制传感器 [23]。虽然这种方法对小规模研究十分有效，但通常难以扩展。此类传感器所需的成本和维护可能使其既昂贵又具有侵扰性，从而降低人们使用它们的意愿。采用商用可穿戴设备的替代方法也无法避免这些挑战，而且这些设备通常被视为小众产品，或在使用一段时间后便被弃用 [25, 76]。这正是智能手机大显身手的地方。在美国，85%的成年人和96%的年轻人拥有智能手机，因此更容易面向更广泛的受众 [17]。移动感知领域的研究揭示了智能手机数据用于活动识别的潜力 [70, 113]。智能手机的广泛普及及其非侵扰性，使其成为传统基于传感器的活动识别的一种极具吸引力的解决方案。然而，我们仍需了解如何利用智能手机中的多种感知模态来进行复杂日常活动识别。此外，复杂日常活动识别模型在不同国家之间的泛化仍是一个研究不足的领域。

识别复杂日常活动十分重要。在活动识别文献中，研究者考虑了多种类型的活动，每种活动处于不同的粒度层级 [33, 99]。粗粒度或简单活动（如步行、坐姿活动或骑行）是可直接通过代理设备（例如惯性传感器单元）测量的重复性单一动作。细粒度复杂活动，即日常生活活动（ADL），建立在简单活动之上，但能够传达更具体的情境信息 [92, 99, 126]。例如，进食、学习、工作和观看电影都包含坐姿活动。这些活动无法仅通过惯性传感器单元进行测量 [9, 12, 73]，而需要一种更为全面的多模态感知方法，以捕捉建立在简单活动之上的广泛情境和行为 [99]。此外，识别此类复杂日常活动还可以：（*i*）以更细粒度的方式追踪个人的数字福祉（例如，细分用于进食、休息、听课和学习的时间，而不仅仅是坐姿活动 [12, 109]）；（*ii*）通过更好地理解用户行为，提供情境感知用户体验和通知（例如，在用户学习或听课时不发送手机通知，以及在用户购物时推荐产品 [75]）；以及（*iii*）实现更好的内容推荐（例如，根据当前的日常活动，如工作、学习或购物，推荐音乐 [125]），因为复杂活动相比简单活动能够提供更多信息并具有更高价值。然而，尽管惯性、位置或WiFi/蓝牙数据已被分别用于活动识别 [92, 99], ，以往研究尚未使用多模态智能手机感知数据对复杂日常活动进行全面研究。

与使用单一模态相比，在机器学习模型中使用多模态智能手机感知数据，能够更全面地呈现复杂日常活动。鉴于新冠疫情导致日常习惯和活动发生重大变化，这一点尤为重要 [112, 135]。为减缓病毒传播而实施的封控措施导致身体活动减少、久坐行为增加，这一现象在年轻成年人中尤为明显。这种转变反映在智能手机使用模式的变化上 [56, 95, 98], ，而在远程/混合工作学习环境中，人们往往会长时间保持久坐，因此这种变化可能会影响基于位置的活动识别方法的有效性。因此，由于久坐行为，惯性传感器和位置传感器作为预测特征的重要性可能会降低。这凸显了纳入细粒度多模态感知特征的必要性，以便通过智能手机准确刻画这些新兴生活方式中的复杂日常活动。然而，目前人们对哪些智能手机感知特征能够系统性地有效刻画不同的复杂日常活动，仍知之甚少。

退后一步，我们也可以考虑“国家”这一维度及其对智能手机使用的影响。不同国家的差异会影响世界不同地区的智能手机使用情况 [66]。例如，在日本，人们可能会觉得在正式餐厅里接听电话有失体统；而在欧洲，人们可能会走出电影院查看手机 [15]。研究表明，与瑞典人相比，日本人通常更不愿在公共交通工具上打电话，或者更普遍地说，不愿在公共场合大声喧哗 [8]。另一项关于智能手机成瘾的研究

该研究调查了24个国家的年轻成年人，发现社会规范和义务的严格程度对智能手机使用有很大影响 [86]。除了人们如何使用电话，先前的研究也探讨了不同国家人群的被动感知行为数据有何差异 [3]。这些国家间的差异构成了一种多样性，这在计算和人工智能研究中正日益受到关注 [27] [1]。从机器学习的角度来看，多样化的系统包含更多信息，因此能够更好地适应各种环境 [43]。更广泛地说，具有多样性感知能力的机器学习旨在通过输入数据、参数和输出等不同组成部分，提升模型的表征能力 [43]。具体而言，具有多样性感知能力的国家层面活动识别，应尝试了解智能手机用户的国家多样性对推断模型性能的影响。然而，目前人们对国家多样性如何影响智能手机感知流程（从收集数据到模型性能）的认识仍然有限，因为先前旨在量化此类影响的研究十分稀少 [52, 70, 89], 。造成这一情况的原因包括但不限于：难以在不同国家按照相同协议开展纵向智能手机感知研究。

我们的工作采用了一系列实验方法（国家特定、国家无关和多国方法，详见表1），以及不同类型的模型（人口层级和混合模型，详见第5节）。借助在多个国家按照相同实验方案收集的丰富多模态智能手机感知数据，我们探讨了三个研究问题：

研究问题1： 复杂日常活动在不同国家中如何表现？哪些智能手机传感特征最有助于区分不同活动？

研究问题2： 通用的多国方法是否适用于复杂日常活动识别？国家特定方法能在多大程度上准确地对国家差异建模？

研究问题3： 复杂日常活动识别模型能否做到国家无关？换言之，在一个或多个国家训练的模型对未见过的国家的泛化效果如何？

为回答上述研究问题，我们提出以下贡献：

贡献 1： 我们研究了一个新颖的智能手机传感器数据集，以及来自五个国家（丹麦、意大利、蒙古、巴拉圭和英国）637 名大学生在四周多时间内提交的超过 21.6 万份自我报告（包括复杂日常活动）。为了表示每份活动自我报告，我们通过处理多模态智能手机传感器数据提取了约 100 个特征（表3）。此外，我们根据参与者回答、活动普遍程度和既有研究，定义了 12 种复杂日常活动类别。类别包括睡眠、学习、进食、观看内容、在线交流与社交媒体使用、上课、工作、休息、阅读、步行、体育和购物。一方面，我们发现，对于某些特定活动（例如睡眠、购物、步行），所有国家中最具信息量的特征都相似。另一方面，对于其他一些活动，最具信息量的特征则因国家而异。不过，有趣的是，地理位置较近的国家之间，这些特征仍大致相似。例如，在欧洲国家中，“运动”活动的一个重要特征是“健康与健身应用”的使用。然而，在蒙古和巴拉圭，这一特征并不突出，因为与身体活动相关的应用使用率较低。“观看内容”活动也呈现出这种差异：在欧洲国家，该活动受到娱乐应用使用的影响，而在另外两个国家则不然。

贡献 2： 我们定义并评估了一项包含 12 个类别的复杂日常活动推断任务，采用国家特定、国家无关和多国方法（表1）。我们还使用人群层面（非个性化）和混合（部分个性化）模型，评估模型个性化如何影响国家内及跨国家的性能。结果显示，通用多国方法直接汇集所有国家的数据（这是许多研究中常见的方法），结合混合模型时 AUROC 达到 0.70。对于这五个国家，国家特定模型的表现最佳，AUROC 分数介于 0.79 至 0.89 之间。这些结果表明，尽管多国家模型使用了更多数据进行训练，却未能整合所有信息以提升性能，这可能是由于不同国家的多样行为产生了平均效应。国家特定方法始终表现更好。

贡献 3： 采用国家无关方法后，我们发现模型难以很好地泛化到其他国家，在总体层面设置下，所有 AUROC 均低于 0.7。采用混合模型后，个性化提升了模型的泛化能力，使 AUROC 分数达到 0.8 以上，但仍未达到国家特定混合模型的相同水平。此外，即使经过部分个性化，我们仍观察到，在欧洲国家训练的模型部署到其他欧洲国家时，其表现优于部署到蒙古或巴拉圭的情况。这表明，除国家多样性之外，行为和技术使用习惯也可能是影响模型在不同国家中性能的因素。鉴于这些发现，我们认为，人机交互和泛在计算研究人员在训练、测试和部署用于理解现实生活中的人类行为及复杂日常活动的系统时，应当意识到机器学习模型所具有的地理敏感性。我们还强调，需要开展更多工作，以应对跨国家多模态移动感知数据集中的领域偏移挑战。

据我们所知，这是首项聚焦于利用多模态智能手机感知数据进行复杂日常活动识别的研究。该研究使用大规模多国数据集，考察了数据的国家层面多样性对复杂活动识别模型的影响，并强调了日常活动识别中的领域偏移相关问题，即使在不同国家采用相同的实验方案收集数据，这些问题依然存在。

本文结构如下：第2节介绍相关工作和背景，随后在第3节介绍数据集。第4节对重要特征进行描述性分析和统计分析。我们在第5节和第6节中定义并评估推断任务。最后，在第7节讨论，并在第8节作结论。

[1]我们承认，文化具有多维性，不同文化之间可能存在张力，同一国家内也可能存在多种文化 [131], 。移动感知、心理学和社会学领域的一些既有研究曾将“文化”用作代理设备，指代数据收集所在的国家 [47, 52, 89, 118]。然而，为保持一致性，本研究使用“国家”（一种更具体的地理区域）作为分析单位，因为国家可能影响手机使用行为和感知数据。我们也仅在适当且涉及区域（即欧洲）时才使用“地理”一词。

表1：本研究中用于描述训练和测试方法及目标类别的术语。

术语	描述
复杂日常活动	基于先前探讨复杂日常活动的研究 [54, 92, 99], 我们将这些定义为贯穿个人日常生活的活动；其本质上较为复杂，并在非瞬时的时间窗口内发生；且具有语义含义和意图，可围绕其构建情境-感知应用程序。
国家特定	该方法使用来自同一单一国家的训练和测试数据。每个国家都有其自身的不利用其他国家的数据的模型。顾名思义，这些模型是特定于各个国家的（例如，在意大利训练并在意大利测试的模型）。
国家无关	这种方法假设数据和模型不受国家影响。因此，训练好的模型可以部署到任何国家，而不受训练所在国家的限制。国家无关阶段有两种： （第一阶段）此阶段使用来自一个国家的训练数据，以及来自另一个国家的测试数据。这对应于已有训练好的机器学习模型，并且我们需要了解它对新国家的泛化能力（例如，在意大利训练并在蒙古测试）。 （第二阶段）此阶段使用来自四个国家的训练数据，以及来自剩余国家的测试数据。这对应于模型已使用来自多个国家的数据完成训练的情形，我们需要了解它对新国家的泛化能力（例如，使用数据训练的模型来自意大利、丹麦、英国和巴拉圭，并在蒙古进行测试）。
多国	这种一刀切的方法使用来自五个国家的训练数据，并在所有国家中测试学习模型。这对应于汇总多国数据以构建一个单一泛化模型的情形。然而，模型通常也是在不考虑国家层面的多样性等方面。

2 背景与相关工作

2.1 移动感知

在以往的研究中，研究人员收集并分析移动感知数据，以了解特定人群的各种属性。根据研究目的，这一目标可归入行为、情境和个体属性识别等宽泛类别 [70]。行为识别旨在广泛了解用户的活动。个体属性识别关注了解人口统计属性（如性别、年龄等）、心理相关属性（如情绪、压力、抑郁等）以及人格。最后，情境识别则用于识别移动设备用户所处的不同情境（如社会情境、位置、环境因素等）。

关于行为识别，有研究旨在捕捉单一复杂活动/行为的二元状态（有时为三种状态），例如进食（如进食正餐与零食 [9], 进食过量、进食不足与按平常进食 [74]）、吸烟（如吸烟或不吸烟 [67]）或饮酒（如饮酒程度 [5, 90], 饮酒或不饮酒 [100]）。另一项研究利用基于音频的导航应用的操作日志来预测其使用情况，并了解影响用户参与度的因素 [60]。随后，在个人层面，MoodScope 系统 [57] 基于用户与电子邮件、电话和短信的互动，以及手机位置和应用使用情况，采用多元线性回归推断智能手机用户的情绪。Servia-Rodriguez 等人 [108] 观察到参与者的日常活动与某些心理变量之间存在相关性。他们训练了一个能够利用智能手机传感器数据预测参与者情绪的深度神经网络。此外，Khwaja 等人 [52] 利用智能手机传感器数据，基于随机森林构建了人格模型。最后，情境识别旨在检测行为和活动周围的情境。 [72] 利用来自瑞士和墨西哥的感知数据，了解大学生进行进食活动时，这些数据与其社会情境之间的关系。更具体地说，他们构建了一个推断模型，用于检测参与者是独自进食还是与他人一起进食。同样，[71] 研究了年轻成年人的智能手机数据，利用加速度计、应用使用、位置、蓝牙和接近度等模态的特征，推断饮酒事件的社会情境。在这种情况下，情境检测包括两个方面：一是基于群体中的人数，二是基于参与者与他们的关系（例如，独处、与另一人一起、与朋友一起、与同事一起）。同样，移动感知研究也试图将利用智能手机传感器感知到的行为和情境作为代理变量，以推断其他情境、心理特质和活动 [26, 49, 70]。

这些研究的一个共同特点是，它们都在真实环境中开展，重点关注两类或三类状态推断，而且感知粒度较粗（即将行为和情境作为因变量的代理）。本文采用了类似的方法，使用在真实生活中采集的数据集和多模态智能手机传感器数据，并将行为和情境作为因变量的代理。不过，本研究的目标属性涉及一个包含12类的日常活动识别问题，与以往研究相比更为复杂且新颖。此外，我们还希望考察在五个国家内部及跨国家的模型性能，并比较是否采用部分个性化的情况。

2.2 活动识别

人体活动识别（HAR）旨在了解人们在特定时刻正在做什么。由智能手机用户活动产生的大规模数据集，在解决这一任务方面具有巨大潜力。这种"数字足迹"已被用于借助信用卡元数据重新识别个人 [30]：研究表明，只需4个数据点就能重新识别90%的人。虽然也可以采用相同的方法利用智能手机感知数据，但我们的重点是在单个时间点进行活动识别，而非使用时间序列进行重新识别。我们将重点关注两类活动识别技术：基于可穿戴设备的技术和基于智能手机的技术 [114]。

2.2.1 基于可穿戴设备的人体活动识别。在基于可穿戴设备的活动识别中，用户佩戴可穿戴加速度计等传感器，通过分析和分类传感器采集的数据来检测活动。例如，在医疗保健领域，基于可穿戴设备的人体活动识别可用于分析步态并预防跌倒，也可用于监测身体活动并观察健康结果 [59]。可穿戴感知的趋势兴起于二十年前，当时依赖定制可穿戴传感器[38, 87], ，而健康研究中令人鼓舞的发现为其提供了支持。随着时间推移，定制传感器被商用健身或活动追踪器取代。遗憾的是，由于定制传感器生产成本高、难以向广大受众分发设备，以及部分用户对其不感兴趣，这些发现很少被应用于真实场景 [25]。因此，大多数使用可穿戴设备的研究仅限于在受控环境中进行实验，或在真实场景中针对较小的人群开展研究。然而，目前许多商用可穿戴设备都已采用能够识别简单活动的基于可穿戴设备的人体活动识别模型。

2.2.2 基于智能手机的活动识别。随着过去二十年智能手机的普及，基于可穿戴设备的人体活动识别所面临的问题得到了解决。Reality Mining [35] 是移动感知领域的一项开创性研究：该研究展示了移动感知数据在自由生活场景中的实用价值。在基于智能手机的活动识别中，人们无需使用可穿戴传感器。相反，该系统依赖一部始终开机且靠近用户的智能手机。智能手机取代了可穿戴设备，因为前者包含加速度计、陀螺仪、全球定位系统、接近度传感器或温度计等多种传感器。然而，智能手机会在多个位置（例如口袋、手中或手提包中）采集数据，这会导致传感器测量值产生偏差，因为这些测量值取决于设备所处的位置 [130]。

无论使用何种设备，以往的大多数活动识别任务都是在实验室或受控环境中进行的，在那里可以准确采集真实值 [113]。这类研究的主要目标是借助精确的真实值和传感器数据收集来提高活动识别模型的准确度（例如，将传感器固定在身体特定部位、通过视频记录真实值等）。然而，这类研究难以扩展，也无法反映参与者的真实行为；对于复杂日常活动而言，这一点尤其明显 [99]。例如，一个人在非受限环境中学习、工作或购物时的行为，无法在实验室中重现。另一方面，也有一些研究在真实场景中开展 [55, 99], ，其真实值和传感器数据收集可能不那么精确，但能够捕捉自然场景中的复杂日常活动。我们的研究与此类似，我们的意图是采取更具探索性的方法，构建能够考虑国家层面多样性的模型，并比较这些模型在不同国家内部及国家之间的性能。

2.3 活动类型

现有研究之间的一个关键差异在于活动的选择。大多数研究致力于识别简单活动。例如，Straczkiewicz 等人 [113] 将活动分为姿势（躺、坐、站立）、移动（步行、爬楼梯、跑步、骑行）和运动（机动活动）等类别。Laput 和 Harrison [54] 将此类活动称为粗粒度活动或全身活动。属于这些类别的活动可以通过一个或多个代理变量（例如惯性传感器单元、位置）直接测量。例如，在考察加速度计时，每项活动在不同轴上都有独特的模式 [33]。然而，这些活动仅占人们日常生活中所进行活动的一小部分 [24, 92, 99]。

请注意，上述描述的一些简单活动通常是更复杂活动的一部分（例如，进食时保持坐姿、购物时步行）。Dernbach 等人 [33] 将复杂活动定义为一系列通常相互重叠的多个动作。他们与 Bao 等人 [7], 一样，采用相同的技术来识别简单活动和复杂活动。由于复杂活动的结构更加复杂，这会导致其识别性能较弱。另一种方法是采用层次化方式考虑复杂活动，通过简单活动的组合来预测更复杂的活动。Huynh 等人 [50] 将用户日常活动描述为简单活动的概率组合。Blanke 等人 [10] 采用自顶向下方法，首先识别简单活动，再据此识别复杂活动。然而，这要求预先定义简单活动以及它们与复杂活动之间的映射关系。一些研究侧重于检测单个复杂活动或特定动作的二元屏幕使用时段。例如，Bites'n'Bits 研究 [9] 考察了正餐与零食之间的情境差异，并提出了一种能够在学生中区分进食事件的分类器。同样，DrinkSense[100] 旨在检测年轻成年人在周末夜晚发生的饮酒事件。遗憾的是，当这些针对特定任务的分类器应用于未经训练的情境时，其性能会很差。

在本研究中，我们关注从五个国家的大学生提供的216K 多份自我报告中归纳出的多数复杂日常活动（12 项中有11 项），以及一项简单活动：步行。在此情境下，借鉴以往对日常生活活动的研究 [92, 99], ，就本文的研究范围而言，我们将复杂活动定义为"贯穿日常生活；本质上复杂，且持续时间并非瞬时；并具有语义含义和意图，可据此构建情境感知应用的活动"。虽然不可能创建一个能够识别所有复杂人类活动的分类器，但我们相信，所提出的分类器涵盖了广泛的常见活动／行为，尤其是年轻成年人中的活动／行为。

2.4 智能手机感知中的多样性意识

包括上述研究在内，智能手机感知领域的研究在人群多样性方面存在不足 [70]。

在国家多样性方面，除少数例外 [52, 108], ，大多数实验都在单个国家开展，极少数则涉及两个国家。这可能会影响研究结果的泛化，因为不同地理区域的智能手机使用情况各不相同，从而可能观察到不同的模式，例如，在性别或年龄范围不同的两个人群中观察到的模式可能有所差异 [31]。Khwaja 等人强调了移动感知中多样性意识的重要性 [52]。此外，在受控环境中开展的实验通常无法纳入许多参与者。虽然这使整个过程更加轻便、易于管理，但也限制了研究结果向更广泛的自由生活人群泛化 [101, 120]。根据 Phan 等人 [89], ，跨国家泛化能力是指研究结果适用于研究对象以外不同群体的程度。

多样性意识和模型泛化是两个重要方面，因为它们将使活动识别系统能够部署，并在不同用户群体和国家中表现良好 [69, 102]。计算机视觉研究反复表明，在性别和肤色等特定属性上存在多样性不足 [28, 51, 93]。在自然语言处理和语音研究中，如果不考虑不同国家的方言，可能会使某些国家的群体被边缘化 [91]。因此，在开发人工智能系统时忽视国家多样性，从长远来看可能会因边缘化某些人群而损害用户利益 [91]。在此背景下，考虑国家层面多样性的智能手机感知研究仍然十分有限 [89]。这可能是由于缺乏大规模数据集、在不同国家开展数据收集存在后勤困难，以及研究耗时耗资。Khwaja 等人 [52] 利用来自五个国家的智能手机传感器数据构建了人格推断模型，并表明此类模型在新国家进行测试时表现良好。据我们所知，他们的研究是最早探究基于智能手机感知的推断模型在不同国家之间泛化能力的研究之一。在我们的工作中，我们专注于利用智能手机感知进行复杂日常活动识别，旨在揭示并考察模型在多国家环境中的行为。

2.5 智能手机使用中的以人为本因素

迄今为止，我们的文献综述主要聚焦于数据收集或目标变量等技术层面。现在，我们将讨论智能手机使用对个人和社会的影响，这一议题受到社会科学多个学科的研究。先前的研究包括对年轻成年人智能手机依赖的研究，该研究发现，问题性智能手机使用因国家和性别而异 [61, 119], ，且社交网络、电子游戏和网上购物等特定活动会助长成瘾 [61, 86]。另一项研究 [96] 总结了青少年和年轻成年人智能手机使用与心理疾病之间相关性的研究发现。过度使用智能手机可能导致情绪困难、冲动性、害羞、低自尊，以及一些医疗问题，例如失眠、焦虑或抑郁。从社会学视角出发，Henriksen 等人 [48] 研究了智能手机如何影响咖啡馆中的互动，并定义了社交性智能手机实践的三个概念。*互动中止*（例如，朋友去洗手间时），这可能导致人们使用智能手机来显得忙碌，或在独处时避免尴尬的情境。*有意屏蔽互动*指暂停正在进行的互动以接听电话或阅读短信，无论是因为紧急情况，还是仅仅因为害怕错过什么。*访问可分享内容*则会让人们共同关注分享的内容（例如图片或短视频），使智能手机发挥增进面对面社交互动而非妨碍互动的作用。Nelson 和 Pieper [81] 指出，智能手机依恋"无意间加剧了绝望感，同时又承诺消除这种绝望"，从而使用户陷入负面循环。

根据 Van Deursen 等人 [119], 的研究，年龄较大的人群较不容易形成智能手机成瘾行为。智能手机虽然常与年轻一代联系在一起，但在年长一代中也逐渐受到欢迎，他们正以富有创意的方式将智能手机融入日常习惯。Miller 等人 [78] 研究了智能手机在九个国家不同社群中所发挥的作用。通过为期 16 个月的民族志研究，他们发现，不同群体的人们会通过应用、个性化设置和通讯，以各自特有的方式掌控自己的智能手机。例如，在爱尔兰，老年人在许多日常活动中都会使用智能手机；在巴西，人们使用即时通讯应用获取健康信息，进而促成了一份关于如何通过此类应用获取健康信息的最佳实践手册。从更广泛的角度来看，智能手机能帮助用户与大家庭成员或远方的朋友保持联系；这一特征在 2020 年全球疫情期间尤为重要。本文试图通过多模态感知数据揭示国家特定的智能手机使用模式。尽管这些见解可能不及实地观察所提供的深入，但它们为未来研究提供了起点。

因此，在综合考虑这些因素的同时，我们旨在研究基于智能手机感知的复杂日常活动识别推断模型，采用国家特定、国家无关和多国家方法，如图 1 所示。

3 数据、特征和目标类别 3.1 数据集信息

为回答我们的研究问题，我们在 2020 年 11 月期间，围绕欧洲项目"WeNet：我们的互联网" [2]，收集了一个关于大学生日常行为的智能手机传感器数据集，历时四周。研究流程总结见一份技术报告 [41]。样本包括本科生和研究生。收集该数据集旨在研究研究参与者的多样性对社交互动和智能手机传感器数据的影响。该数据集包含来自居住在五个国家（按参与者人数排序）的 637 名大学生所报告的逾 21.6 万项活动：意大利、蒙古、英国、丹麦和巴拉圭。所有数据均通过使用同一款移动应用的安卓智能手机收集。表2显示了各国参与者的分布情况。

[2]该数据集计划在项目结束后面向研究用途发布，并遵守欧盟境内外有关数据收集规程的所有法规。因此，未来的数据集访问计划将公布于项目网站：https://www.internetofus.eu/

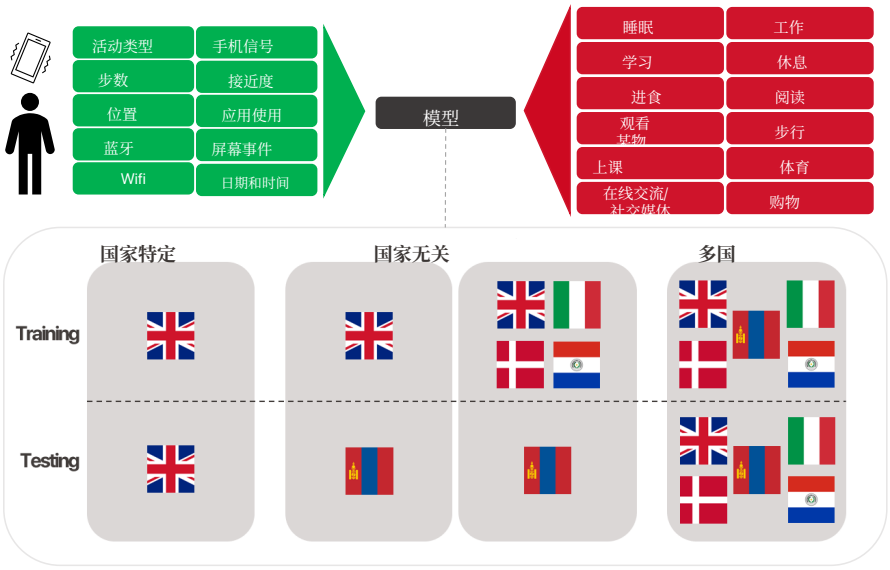


图 1：研究的高层概览。本研究采用连续感知和交互感知模态，并使用不同的方法（国家特定、国家无关和多国）来推断复杂日常活动。

此外，数据收集所采用的协议符合欧盟《通用数据保护条例》（GDPR）以及各非欧盟国家的相关规定。此外，每所参与研究的大学均分别取得了伦理审查委员会（或类似机构）的书面批准。

数据收集的第一阶段获取了关于参与者、其习惯、社会关系、个人实践（例如体育活动、休闲活动）以及技能（个人技能和人际交往技能）的问卷数据。这些数据旨在捕捉多样性的不同方面，包括人口统计学等可观察特征，以及人格、特质、技能、价值观和关系等较不易观察的方面 [103]。第二阶段通过一款智能手机应用程序收集数据。参与者每天多次填写时间日记。每天开始时，参与者需要回答有关睡眠质量和预期的问题。一天结束时，他们必须报告当天的情况。每小时，他们都必须自我报告正在进行的活动（当前活动，从包含 34 项活动的下拉列表中选择）、位置（包含 26 个语义位置的列表）、社会情境（包含 8 个社会情境的列表）和情绪（效价的测量方式与 [57] 类似，采用五点评分量表）。该应用程序持续收集来自三十多种智能手机传感器的数据，这些数据可分为两类 [70]：连续感知模态，例如简单活动类型（使用惯性传感器和位置，并通过 Google 活动识别 API [44] 得出）、步数、位置、Wi-Fi、蓝牙、手机信号、电池和接近度；以及交互感知模态，例如应用程序使用时长、屏幕使用次数和时长、通知点击行为以及用户在场时长。

3.2 特征推导

数据集的格式选择对研究的其余部分至关重要。以活动或事件为中心的表格数据集能够实现先前文献中讨论的大量传感器特定特征的手工构建 [16, 70, 74, 100, 108]。这使得后续可以采用传统的机器学习方法。然而，时序数据集主要依赖以时间序列形式呈现的原始传感器测量值（即典型活动识别中的原始加速度计和陀螺仪数据）。这种方法能够让深度学习方法自动提取并学习相关的高层特征。既有研究 [4, 137] 表明，使用深度学习技术可以获得性能更优的人类活动识别分类器。这些研究通常包括更容易通过惯性传感器检测的简单活动，而不是我们所关注的更复杂活动。在远程学习或工作环境中，这一点尤为重要，因为许多活动都是在家中进行的。因此，我们选择使用包含下文所述异构手工特征的表格数据集进行分析。

我们将所有传感器测量值与自我报告汇总起来，利用智能手机数据生成特征。我们采用了与以往事件级推断研究类似的基于时间窗口的方法 [70, 108, 114]。因此，我们取每条自我报告前后各10分钟的数据，并汇总相应20分钟区间内的传感器数据 [3]。传统的简单活动识别以及基于惯性传感器的识别，会尝试利用深度学习和较短的时间窗口捕捉重复发生的片段；但这种方法不适用于此处，因为我们尝试捕捉一组持续时间较长的非重复性活动。此外，我们将智能手机感知到的行为和情境视为目标活动的代理设备，

[3]我们针对 5 分钟到 25 分钟之间的不同时间窗口进行了实验。我们没有采用超过 25 分钟的时间窗口，因为这会导致传感器数据片段重叠，从而造成数据点之间的数据泄漏。在所测试的时间窗口中，20 分钟的窗口表现最佳。为简洁起见，我们仅展示采用 20 分钟窗口的结果。较短的窗口可能无法取得理想效果，因为它们无法捕获足够的情境信息来进行推断。以往研究也表明，较长的时间窗口可能适合检测二元活动[6, 9, 71]

表2：数据收集参与者概况。国家按照参与者人数排序。

大学	国家	参与者	μ 年龄（σ）	% 女性	自我报告数量
特伦托大学	意大利	259	24.1（3.3）	58	116,170
蒙古国立大学	蒙古	224	22.0 (3.1)	65	65,387
伦敦政治经济学院与政治学	UK	86	26.6 (5.0)	66	20,238
天主教大学"圣母升天教堂"	巴拉圭	42	25.3 (5.1)	60	6,998
奥尔堡大学	丹麦	26	30.2 (6.3)	58	7,461
总计/均值		637	24.0 (4.3)	62	216,254

这与以往的普适计算研究类似 [67, 74, 108]。因此，利用各种感知模态生成的相应特征如表3所示。有关各感知模态预处理方式的更多详情，见 [69]。除了传感器数据特征外，我们还添加了一个特征，用于描述活动发生时一天中的时段，以及一个周末指示变量。尽管文献中对于如何将一天划分为早晨、下午、傍晚和夜晚尚无共识 [82, 84, 121, 122, 124],，我们将一天划分为五个时段：早晨为上午6点至10点，中午为上午10点至下午2点，下午为下午2点至6点，傍晚为下午6点至10点，夜晚为晚上10点至次日早上6点，并将其作为另一个可用于训练机器学习模型的特征。

3.3 确定目标类别

参与者需要每小时进行自我报告，从预先设定的34项活动列表中选择一项，记录当时正在做什么。这些项目是根据先前的研究制定的 [40, 134]。观察各国的分布情况（图2），可以很快发现这些活动的分布极不平衡。数据收集期间的远程工作/学习限制是造成这种不平衡的原因之一，因为如果移动不受限制，旅行、步行或购物等活动会更常见。仔细查看活动列表会发现，有些类别的语义含义过于宽泛。因此，与先前基于多个方面缩减活动列表的研究类似 [54],，我们将原始活动列表缩减为12个类别，以捕捉人们日常生活中足够常见的复杂日常活动，尤其是在远程工作/学习情境下。例如，“爱好”可能包括弹钢琴或绘画，但这两者所涉及的智能手机使用并不相同，而且都不够常见。同样，“社交生活”的范围也过于宽泛，因为人们可能身处酒吧、餐馆或公园。此外，为缓解类别不平衡问题，我们决定筛选目标类别。首先，合并语义含义相近的类别，例如进食与烹饪、社交媒体与网上聊天。其次，移除代表宽泛活动的类别，例如个人护理、家务、游戏和爱好。最后，移除在所有国家都没有足够数据的类别，例如听音乐、电影、戏剧、音乐会和业余学习。“没什么特别的”或“其他”等填充类别也被移除。经过筛选，目标类别减少至12个，其更新后的分布如图3所示。这些类别涵盖了人们日常生活中具有复杂性质且具有语义含义的活动，围绕这些活动可以构建情境感知应用。此外，所选活动也与先前研究复杂日常活动识别的工作相一致 [99]。

4 不同国家如何表现活动，哪些智能手机特征最具区分性？（研究问题1）

为了了解各国活动的分布，并确定特征对目标的影响，本节对数据集进行了描述性统计分析，从而阐明**研究问题1**。

4.1 活动的每小时分布

我们考察的活动似乎都发生在不同的时间：人们通常夜晚睡觉，白天工作，并在中午和傍晚进食。然而，并非所有人的日程安排都一样，不同国家之间尤其如此 [37, 39]。我们在图 4 中报告了每个目标类别在一天中不同时段的密度函数。每个图表中，x 轴表示一天中的时段，y 轴表示各项活动的密度。需要特别指出的是，虽然大多数活动都是在参与者进行活动时报告的，但对于睡眠，参与者是在醒来后、仍躺在床上时报告该活动的，这意味着该活动的峰值也可以解释为“醒来”。后来，在研究后访谈中，所有国家的许多参与者都证实了这一点。这也使得一天中的时间在推断睡眠活动时参考价值较低。

初步观察分布可以发现一些“预期之中”的模式，例如夜间睡眠时段出现一个小高峰，午餐和晚餐时段则出现高峰。值得注意的是，来自巴拉圭的参与者往往比其他人睡得少，这反映出他们一天中开始工作和休息的时间更早。在线交流和社交媒体使用在中午左右较为频繁，恰逢课间休息和午餐时间，之后在一天结束前出现一个高峰。这与以往研究的发现一致：根据位置情境和一天中的时段，某些社交媒体应用（即 Twitter）的使用情况可能有所不同 [32]。此外，我们还观察到用户报告的社交媒体和在线交流应用使用模式存在国家间的时段差异。例如，中午至下午 6 点期间，意大利、巴拉圭和丹麦的这类应用使用量有所下降，而英国则没有出现这种模式。以往研究也探讨过社交

表3：使用时间窗口围绕活动自我报告进行汇总后，从原始感知数据中提取的108个特征的摘要。

模态	对应的特征及描述
位置	回转半径、行进距离、平均海拔的计算方式与先前研究类似 [16]
蓝牙 [LE，普通]	设备数量（发现的唯一设备总数）、接收信号强度指示的均值/标准差/最小值/最大值（接收信号强度指示——用于衡量设备距离远近） [100]
WiFi	已连接到网络指示器、设备数量（发现的唯一设备总数）、接收信号强度指示的均值/标准差/最小值/最大值 [100]
蜂窝网络 [全球移动通信系统，宽带码分多址，LTE]	设备数量（发现的唯一设备总数），手机信号强度的均值/标准差/最小值/最大值 [100]
通知	已发布的通知（传送到电话的通知数量）、已移除的通知（数量为参与者移除的通知数）——这些特征分别在计入和不计入重复项的情况下计算。[57]
接近度	接近度值的均值/标准差/最小值/最大值 [6]
活动	进行以下简单活动的时间：静止、乘坐_车辆、骑_自行车、步行_、跑步、倾斜、步行，其他（使用 Google 活动识别 API [44]推导）
步数	步数计数器（根据上次电话开机以来的总步数推算步数），检测到的步数（根据步数推算每检测到新的一步就触发的事件） [29]
屏幕事件	触摸事件（电话触摸事件的数量）、用户在场时长（通过 Android API 推算，该 API 可指示某人是否正在使用电话）、屏幕使用时段数量（一个屏幕使用时段从电话屏幕开启到屏幕关闭为止）、时段时长的均值/最小值/最大值/标准差（一个时间窗口内可能包含多个屏幕使用时段），总时长（时间窗口内屏幕开启的总时长） [6, 57, 74]
应用事件	根据 Google Play 商店分类统计的各类别应用使用时长 [57, 100]：动作、冒险、街机、艺术与设计、汽车与交通工具、美容、棋盘游戏、图书与参考资料、商业、纸牌游戏、赌场游戏、休闲、漫画、通讯、约会、教育、娱乐、金融、美食与饮品、健康与健身、家居、生活方式、地图与导航、医疗、音乐、新闻与杂志、育儿、个性化、摄影、效率、益智游戏、竞速、角色扮演游戏-玩乐、购物、模拟、社交、体育、策略、工具、旅行、问答游戏、视频播放器与编辑器、天气，word
时间与日期	一天中的时段、一天中的时段（早晨、中午、下午、傍晚、夜晚）、周末指示变量（工作日或周末） [9, 74]

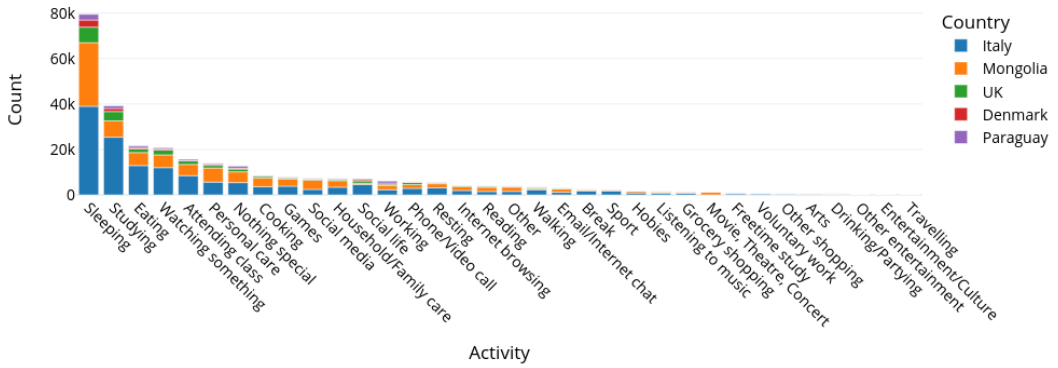


图 2：在进行任何筛选或合并之前，目标类别的原始分布 **e.**

媒体应用使用及其使用差异，尤其是国家间的差异。根据这些研究，此类使用差异可能源于各国的文化特征，也可能与人们使用不同应用的动机有关 [2, 58]。大多数休闲活动（阅读、购物、运动、观看内容）都发生在一天结束前后，正值学生结束课程之时。

另一个呈现出明显跨国差异的活动是“进食”。我们可以观察到，意大利人往往比其他人更晚进食，这暗示了他们的地中海习俗 [117]。意大利在午餐和晚餐时段也呈现出两个明显的峰值，而两顿正餐之间则出现了急剧下降。其他国家的这一低谷不那么明显，表明正餐分布在不同时间段，间隔更为分散。此外，除蒙古以外，所有国家的晚餐高峰均出现在下午 6 点或之后，而蒙古则早于下午 6 点。这些发现表明，一天中的时段可以指示人们是否正在进食——但在意大利、蒙古和其他国家之间略有不同。事实上，先前使用移动传感器开展饮食行为研究的研究表明，一天中的时段是预测与进食相关方面的重要特征 [9, 74]。此外，先前的研究还指出，不同国家之间的用餐时间、频率和份量也可能存在差异 [21],，即使是在欧洲内部也是如此 [111]。最后，“步行”活动在不同国家之间的分布更不相似，或

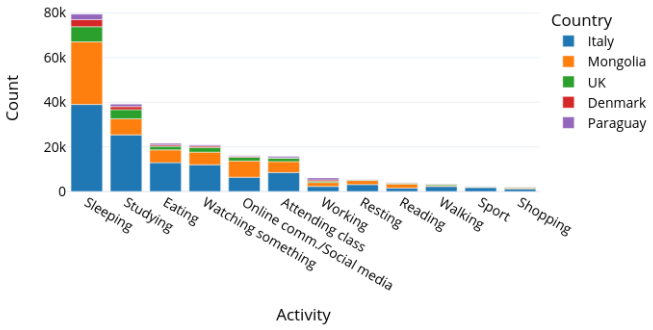


图3：移除语义范围过广或缺少数据的类别后，目标类别的分布。

少相似。事实上，Althoff 等人开展的一项基于智能手机的活动追踪研究 [3] 提到，意大利、英国、丹麦和蒙古人群的平均步行步数大致处于同一水平（即每日步数约为 5000—6000 步）。

4.2 特征的统计分析

为了解每种智能手机传感特征在区分各目标活动与其他活动时的重要性，我们在表4中报告了每个国家、每项活动对应的前三个特征及其方差分析（方差分析）F值 [53] 。我们的目标是确定定义某项活动的特征，以及这些特征在不同国家之间的差异。我们单独考察每个国家—活动组合，以找出在二元情境下会影响分类任务的特征（即判断参与者是否正在睡眠、学习、进食等）。

在大多数情况下，同一活动在不同国家呈现出的特征各不相同，凸显了数据集的多样性以及各国的文化差异或习惯。例如，在学习时，屏幕使用片段相关的特征在英国、意大利和丹麦占主导地位，而一天中的时段则出现在意大利、蒙古和巴拉圭。这可能意味着，欧洲学生在学习时使用电话的频率高于（或低于）来自巴拉圭或蒙古的学生。在"观看内容"这一活动中也能看到这种分化：该活动受到欧洲娱乐应用使用情况的影响，但在巴拉圭或蒙古则并非如此。这种影响可能是由于在后两个国家，被归类为娱乐应用的流媒体服务并不普及，参与者可能会依赖其他替代方案。事实上，先前的研究已经考察了不同国家之间流媒体服务的使用差异，突出了使用比例 [62] 以及其与收入水平之间的关系 [83]。另一方面，也可能是学生通过并非电话的媒介观看内容。事实上，研究表明，在美国，18至29岁的年轻成年人相比电视，会更多地使用在线媒体流媒体服务 [18]。然而，在文化、收入水平和互联网质量存在差异的不同国家之间，是否也保持相似的比例，仍然是一个有待回答的问题。尽管尚无定论，这些因素可能正是娱乐应用无法成为蒙古和巴拉圭"观看内容"活动指示特征的原因而蒙古和巴拉圭正是本研究中的非欧洲国家。

对于某些活动，排名前三的特征是该活动本身固有的。例如，在意大利，"阅读"对应的特征包括书籍、漫画、报纸和杂志等阅读应用。其他国家则没有这种现象。"体育"活动也有类似观察：在欧洲国家，健康与健身应用是决定性特征之一。这种现象可能与参与者使用智能手机应用追踪自己的锻炼有关。

"步行"活动在五个国家中的特征几乎完全相同：检测到的步数，以及通过 Google 活动识别 API 检测到的非步行或步行活动。这种同质性源于该活动的性质——步行被视为一种简单的活动。这也解释了为什么购物与步行具有一些相同的特征，因为参与者在购物时也会步行。总而言之，在大多数情况下，考察同一活动时，每个国家都有不同的定义性特征。对于某些活动而言，所发现的特征是该活动固有的，通常表现为应用类别。最后，值得一提的是，一天中的时段是一项重要特征，这与图 4 中观察到的情况相吻合——所有活动在一天中的发生频率并不相同。

最后，值得注意的是，我们可以预期，随着不同国家技术使用方式和人们习惯的变化，一些信息量很高的特征也会随时间发生变化 [1, 128]。例如，某些国家缺乏流媒体服务使用的一个原因，是围绕非法下载内容使用的相关法律不完善（例如，德国对不使用非法下载内容有严格的法律规定 [97]）。各国法律的变化可能会改变年轻成年人的行为。此外，互联网价格也可能影响流媒体服务的使用。基于带宽且价格低廉的互联网在发达国家很常见，但在亚洲、非洲和南美洲的发展中国家却并非如此；这些地区的互联网使用成本较高，因此会降低人们使用流媒体服务的意愿。此外，收入水平也可能显著影响所捕获的特征。例如，随着收入水平提高（通常发生在一个国家的国内生产总值发生变化时），年轻成年人可能会更多地使用可穿戴设备进行健身追踪，从而推动他们在手机上使用健康与健身应用。可能影响所捕获行为的另一个因素是天气状况。本研究提到的五个国家都经历不同的季节，因为它们都在一定程度上远离赤道。因此，我们可以预期，不同季节中的特征会发生变化。局限性部分将进一步讨论这一点。

5 基于机器学习的推断：实验设置、模型和性能指标

本研究旨在对智能手机感知数据进行多分类推断，以预测参与者在特定时间的活动。输入空间由前文提及的表格数据集中的特征构成。我们研究了该问题的三种方法，如图 1 所示，范围从国家特定方法到多国家方法。

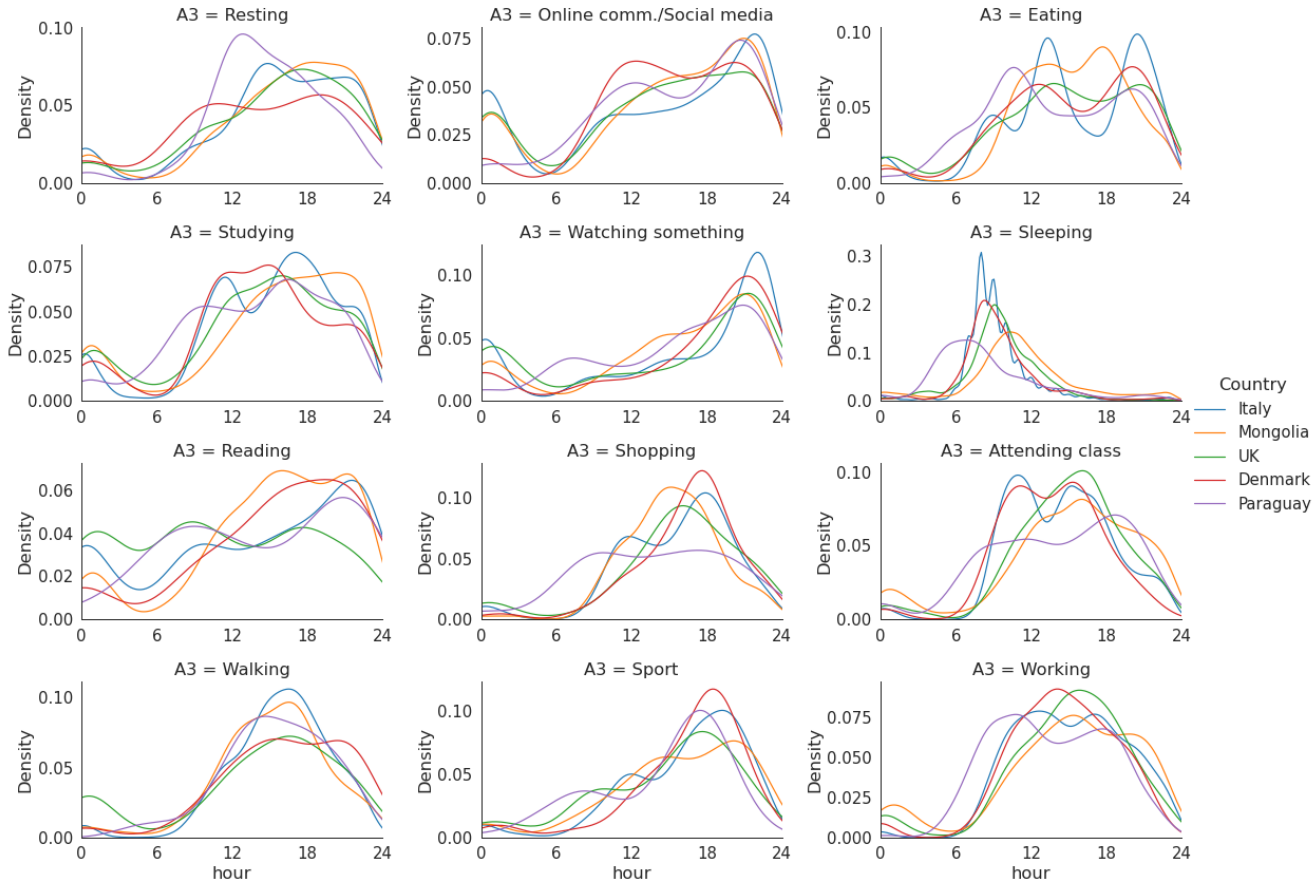


图 4：各国目标类别的密度函数随一天中时刻的变化

y.

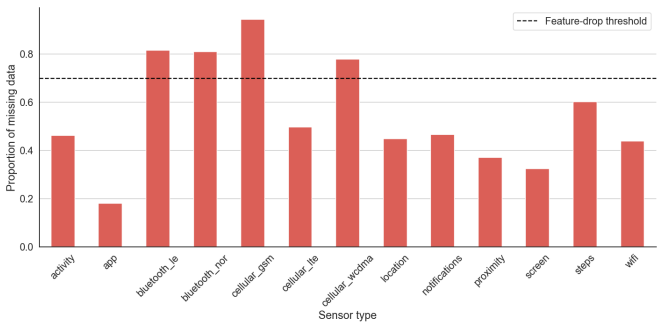


图 5：各传感器类型缺失数据的比例。

5.1 数据插补

为进行推断，准备数据集的第一步是数据插补。在智能手机感知的情境下，数据缺失可能有多种原因 [6, 68, 108]：设备处于低功耗模式、传感器故障，或参与者授予的权限不足。在我们使用的数据集中，我们注意到大多数传感器都存在一些缺失值（见图 5）。例如，超过 90% 的 GSM 蜂窝传感器数据不可用，可能是因为设备处于飞行模式、传感器故障，或电话主要使用 LTE 信号运行。为处理缺失值，我们决定删除缺失数据超过 70% 的传感器特征（参见图 5 中的虚线），这与先前的研究类似 [100]。对于剩余特征，我们针对每个国家分别使用 k近邻（kNN）插补 [133]，根据邻近样本推断缺失信息 [4]。

5.2 模型与性能指标

为开展所有实验，我们使用了 scikit-learn [88]、Keras [22] 框架以及 Python。我们首先训练了以下两个基线模型：一个始终预测出现频率最高的标签，另一个根据类别分布随机预测目标。这将帮助我们了解训练后的模型是否优于随机猜测。实验

[4]除 kNN 外，我们还尝试了均值插补、基于用户的均值插补、众数插补和前向填充（LOCF）插补。不过，使用 kNN 进行推断时，我们得到了最佳结果。此外，在使用被动感知的研究中，kNN 也很常见 [94, 129, 132, 136]。因此，我们只报告了使用 kNN 得到的结果。

表4：每个目标活动和每个国家的方差分析 F 值（F）及 p 值 < 0.05。列表中的第一个特征为最佳特征。仅可在同一活动和国家内比较 F 值。

	意大利		蒙古		UK		丹麦		巴拉圭	
	特征	*F*	特征	*F*	特征	*F*	特征	*F*	特征	*F*
睡眠	工具应用	5423	时段	6623	时段	1632	时段	354	未找到应用	510
	时段	4439	未找到应用	3595	单次最长屏幕使用时长	603	单次最长屏幕使用时长	249	去重后移除的通知	348
	单次最长屏幕使用时长	2498	去重后移除的通知	1052	每次屏幕使用时长	534	每次屏幕使用时长	156	去重后发布的通知	289
学习	单次最长屏幕使用时长	1447	时段	683	总屏幕使用时间	446	单次最长屏幕使用时长	241	视频播放器和编辑器应用	147
	总屏幕使用时间	1378	去重后移除的通知	220	单次最长屏幕使用时长	396	总屏幕使用时间	225	未找到应用	84
	时段	1146	摄影应用	178	每次屏幕使用时长	247	周末	154	时段	43
进食	时段	271	时段	518	时段	61	接近度标准差	38	未找到应用	37
	工具应用	98	未找到应用	180	未找到应用	26	接近度最大值	29	Wi-Fi 平均 RSSI	23
	未找到应用	61	静止活动	72	视频播放器和编辑器应用	23	通信类应用	18	Wi-Fi 最大 RSSI	21
观看某物	娱乐应用	715	时段	326	视频播放器和编辑器应用	397	娱乐应用	151	Wi-Fi 平均 RSSI	51
	未找到应用	426	未找到应用	325	Wi-Fi RSSI 标准差	85	未找到应用	59	生活方式应用	38
	周末	334	Wi-Fi设备数量	217	娱乐应用	66	已发布通知	58	周末	29
在线交流/社交媒体	社交类应用	1381	触摸事件	503	Wi-Fi设备数量	112	工具应用	64	工具应用	95
	总屏幕使用时间	565	总屏幕使用时间	355	已连接Wi-Fi	93	因果类应用	58	接近度最大值	58
	单次最长屏幕使用时长	473	未找到应用	354	总屏幕使用时间	92	总屏幕使用时间	42	平均邻近度	48
上课	周末	3167	时段	455	周末	357	未找到应用	119	去重后发布的通知	148
	屏幕使用次数	745	周末	289	时段	260	已发布通知	104	周末	112
	工具应用	476	未找到应用	251	单次最长屏幕使用时长	70	单次最长屏幕使用时长	37	总屏幕使用时间	87
工作	检测到的步数	271	Wi-Fi平均接收信号强度	1049	每次屏幕使用时长	143	平均邻近度	305	车内活动	441
	每次屏幕使用时长	210	Wi-Fi 最大接收信号强度	848	平均邻近度	129	接近度最大值	304	Wi-Fi设备数量	226
	屏幕使用次数	206	Wi-Fi 最小接收信号强度指示	633	单次最长屏幕使用时长	124	接近度标准差	292	步行活动	163
休息	时段	337	时段	191	医疗应用	374	已发布通知	22	摄影应用	145
	工具应用	117	总屏幕使用时间	89	街机应用	72	未找到应用	16	知识问答应用	64
	教育应用	66	单次最长屏幕使用时长	75	时段	55	触摸事件	14	地图与导航应用	23
阅读	图书与参考应用	955	未找到应用	167	未找到应用	215	最低蜂窝 LTE	252	冒险应用	21
	漫画应用	93	触摸事件	122	Wi-Fi RSSI 标准差	109	工具应用	83	漫画应用	16
	新闻与杂志应用	93	时段	121	Wi-Fi 最大接收信号强度	77	位置海拔	76	位置海拔	6
步行	步行活动	3518	步行活动	1582	检测到的步数	376	检测到的步数	285	步行活动	25
	步行活动	3497	步行活动	1579	步数计数器	314	步行活动	101	步行活动	25
	检测到的步数	3374	检测到的步数	1009	步行活动	232	步行活动	101	位置回转半径	23
运动	健康与健身应用	502	时段	33	健康与健身应用	931	健康与健身应用	1248	Wi-Fi 最大接收信号强度	50
	时段	233	Wi-Fi设备数量	32	最小接近度	52	已移除通知	72	接近度标准差	41
	已发布通知	132	wifi_min_rssi	23	day_period	40	day_period	34	wifi_mean_rssi	41
购物	steps_detected	283	activity_onfoot	1270	时段	74	步行活动	132	天气应用	86
	步行活动	267	步行活动	1269	用户在场时间	41	步行活动	131	汽车与车辆应用	84
	步行活动	265	检测到的步数	504	屏幕使用次数	38	检测到的步数	55	步行活动	79

采用了以下模型类型：随机森林分类器 [14]（RF）、使用决策树分类器 [46], 的 AdaBoost，以及多层感知机神经网络（MLP） [123] [5]。前两者能够自然地处理类别不平衡，而随机森林还便于解释结果。每项实验均重复进行十次，以将随机性的影响纳入考虑。对于每种实验设置，我们报告了以下指标在十次运行中的均值和标准差：F1 分数 [107], 以及受试者工作特征曲线下面积（AUROC） [106]。尽管我们计算了模型的准确度，而且准确度易于解释，但在不平衡数据场景下，它可能无法呈现真实情况。因此，我们未将其纳入结果。加权宏平均 F1 分数会计算每个类别的指标，并根据其支持度进行平均，从而得到一个能够考虑标签不平衡的指标。此外，如果某个类别包含大量假阳性，该指标会受到显著影响。较低的 F1 分数可能意味着分类器难以识别稀有目标类别。AUROC 分数衡量模型区分各项活动的能力。它可以理解为不同阈值下 F1 分数的平均值。我们还使用了加权宏平均版本，以考虑标签不平衡。

[5]除报告的模型外，我们最初还尝试了梯度提升和 XGBoost 等其他模型类型。鉴于这些模型的性能和篇幅限制，未报告其结果。所有这些模型类型通常用于采用表格格式的小型移动感知数据集。 [9, 74, 77]

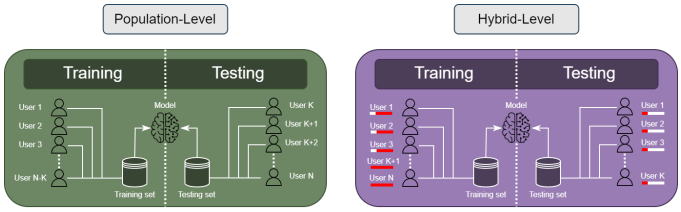


图 6：国家特定、国家无关和多国方法所采用的个性化程度。人口层级对应于没有个性化的模型。混合对应于部分个性化的模型。

接下来，我们考察国家特定、国家无关和多国方法的结果 [52]。最后，对于这三种方法，我们考察人口层级模型和混合模型，分别对应于没有个性化和部分个性化，这与 [57, 68, 69] 类似（训练和测试始终按 70:30 的比例划分）：

- **人口层级** 模型，也称为国家特定和多国方法中的留出 k 名参与者，以及国家无关方法中的留出 k 个国家：训练集（≈70%）和测试集（≈30%）中的参与者互不重叠。数据划分采用分层方式，即每次划分都保留各类别样本所占的百分比

。这代表这样一种情况：模型在部分人群上进行了训练，随后一组新的参与者加入了运行该模型的系统并开始使用它。

– 在国家特定方法中，这意味着训练集和测试集中的参与者彼此不重叠，且所有人都来自同一个国家。例如，使用一组来自意大利的参与者进行训练，并使用另一组未出现在训练集中的意大利参与者进行测试。
– 在国家无关方法中，这意味着训练集来自一个国家（第一阶段）或四个国家（第二阶段），而测试集来自训练期间未见过的国家。例如，对于第一阶段——使用一组来自意大利的参与者进行训练，并使用一组来自蒙古的参与者进行测试；对于第二阶段——使用来自意大利、丹麦、英国和蒙古的参与者进行训练，并使用一组来自巴拉圭的参与者进行测试。
– 在多国方法中，这意味着训练集和测试集中的参与者彼此不重叠，不考虑国家信息。即使数据来自多个国家，这也是训练模型的典型方式 [108]。例如，使用来自五个国家的参与者进行训练，并使用来自这五个国家且未出现在训练集中的参与者进行测试。

• **混合** 模型，也称为留出 k 个样本：训练集和测试集中的参与者集合并不互斥。测试集中部分参与者的数据（≈70%）被用于训练模型。测试时使用这些参与者的其余数据（≈30%）。这代表了这样一种情况：模型在人群数据上训练完成后，训练中使用过其数据的同一批参与者继续使用该模型。因此，模型是部分个性化的。

– 在国家特定设置中，这意味着测试集里某个国家参与者的部分数据也可能出现在训练集中。这代表了在该国内考察个性化的情形。例如，使用意大利的一组参与者进行训练，并使用意大利另一组参与者的数据进行测试，其中 70% 的数据也被用于训练集。其余 30% 的数据则用于测试集。

– 在国家无关设置中，这意味着训练集来自一个或多个国家，而测试集来自另一个国家；该国参与者过去的数据中有一定比例（70%）也被纳入训练集。这代表了模型部署到一个新国家时考察个性化的情形。例如，第一阶段——使用意大利的一组参与者进行训练，并使用蒙古的一组参与者进行测试，其中 70% 的数据也被用于训练集。其余 30% 的数据则用于测试集；第二阶段——使用意大利、丹麦、英国和蒙古的参与者进行训练，并使用巴拉圭的一组参与者进行测试，其中 70% 的数据也被用于训练集。其余 30% 的数据则用于测试集。

– 在多国设置中，这意味着训练和测试的参与者并非互不重叠，且不考虑国家信息。即使数据来自多个国家，这也是对模型进行部分个性化的典型方式。例如，使用来自五个国家所有参与者的数据进行训练，并使用来自五个国家所有参与者的数据进行测试，其中70%的数据也用于训练集，其余30%的数据则用于测试集。

6 推断结果

本节介绍实验结果。首先，我们讨论国家特定方法和多国方法的结果，以阐明**研究问题2**。随后，我们讨论国家无关方法，并针对模型泛化回答**研究问题3** 。

6.1 国家特定与多国方法（研究问题2）

国家特定方法。 我们将国家特定方法视为基础设置：它确实利用了国家层面的多样性，为每个国家分别建立模型，各国模型彼此独立。表5总结了采用国家特定方法开展的实验结果。在总体层面设置中，三种模型的表现大致相当，但根据 F1分数和 AUROC 分数，随机森林通常表现更好。在混合模型中，随机森林在五个国家的表现均最佳，AUROC 分数介于 0.79–0.89 之间；蒙古的分数最低，丹麦的分数最高。与总体层面模型相比，混合模型的性能显著提升，体现了在各国内进行个性化处理的效果。这些结果表明，在部分个性化设置下应用随机森林模型，能够以良好的性能利用被动感知数据识别复杂日常活动。鉴于这一结论，尽管后续章节中我们获得了所有模型类型的结果，但只会展示随机森林模型的结果。

多国方法。 这种方法旨在构建一个通用的多国模型或一刀切模型，以期捕捉所有国家的多样性。五个国家都出现在训练集和测试集中。因此，我们将数据集中的所有参与者一视同仁，不考虑他们所属的国家，这类似于忽略国家层面多样性的实验。由此，我们可以在这种情境下考察多国方法中的人口层级模型和混合模型。此外，我们使用来自五个国家且表示不均衡的数据集评估模型（多国不含下采样 — MC w/o DS），并使用来自五个国家且表示均衡的数据集进行评估；后者通过从数据较多的国家随机下采样，使其数据量与自我报告数量最少的国家（即巴拉圭）相等（多国含下采样 — MC w/ DS）。结果如图7所示，并与国家特定结果进行了比较。MC w/o DS 的 AUROC 为 0.71，而 MC w/ DS 的 AUROC 为 0.68，这表明使用原始数据分布进行训练的效果更好。原因实际上可能是数据越多，性能越好。使用下采样数据进行训练的预期目的是让每个国家得到同等重视，并期望模型在所有国家都能表现良好。然而，结果表明事实并非如此。

这些结果为我们的**研究问题2**提供了启示：仅使用被动智能手机感知数据来训练用于复杂活动识别的多国家模型颇具挑战（混合模型的 AUROC 为 0.709）。其性能并不优于特定国家的方法（AUROC 范围为 0.791–0.894）。这可能源于各国家和类别之间的数据不平衡，或数据集收集时的情境

表5：分别针对每个国家，使用三种不同模型（以及两个基线）经过十次迭代计算得出的推理 F1 分数和 AUROC 分数的均值（$\bar{s}$）和标准差（s_σ）。结果以 $\bar{s}$（s_σ） 的形式呈现，其中 s表示两项指标中的任意一项。

	基线 I		基线 II		随机森林		AdaBoost		MLP	
					人口层级					
	F1	*AUROC*	*F1*	*AUROC*	*F1*	*AUROC*	*F1*	*AUROC*	*F1*	*AUROC*
意大利	0.17 (0.000)	0.50 (0.000)	0.19 (0.001)	0.50 (0.001)	0.41 (0.001)	0.71 (0.001)	0.39 (0.000)	0.71 (0.000)	0.38 (0.002)	0.68 (0.002)
蒙古	0.26 (0.000)	0.50 (0.000)	0.23 (0.001)	0.50 (0.001)	0.33 (0.002)	0.62 (0.001)	0.33 (0.000)	0.63 (0.000)	0.34 (0.003)	0.61 (0.004)
UK	0.17 (0.000)	0.50 (0.000)	0.18 (0.002)	0.50 (0.001)	0.32 (0.004)	0.63 (0.003)	0.31 (0.000)	0.59 (0.000)	0.22 (0.006)	0.56 (0.003)
丹麦	0.25 (0.000)	0.50 (0.000)	0.24 (0.006)	0.49 (0.003)	0.32 (0.008)	0.61 (0.006)	0.34 (0.000)	0.57 (0.000)	0.25 (0.008)	0.57 (0.006)
巴拉圭	0.19 (0.000)	0.50 (0.000)	0.19 (0.006)	0.49 (0.002)	0.30 (0.004)	0.59 (0.003)	0.28 (0.000)	0.56 (0.000)	0.31 (0.009)	0.58 (0.004)
					混合					
	F1	*AUROC*	*F1*	*AUROC*	*F1*	*AUROC*	*F1*	*AUROC*	*F1*	*AUROC*
意大利	0.17 (0.000)	0.50 (0.000)	0.19 (0.001)	0.50 (0.001)	0.63 (0.001)	0.87 (0.001)	0.40 (0.000)	0.73 (0.000)	0.51 (0.002)	0.81 (0.000)
蒙古	0.26 (0.000)	0.50 (0.000)	0.23 (0.002)	0.50 (0.001)	0.51 (0.001)	0.79 (0.001)	0.34 (0.000)	0.66 (0.000)	0.45 (0.002)	0.75 (0.002)
UK	0.17 (0.000)	0.50 (0.000)	0.19 (0.003)	0.50 (0.001)	0.66 (0.001)	0.88 (0.006)	0.34 (0.000)	0.68 (0.000)	0.58 (0.003)	0.83 (0.002)
丹麦	0.25 (0.000)	0.50 (0.000)	0.24 (0.003)	0.50 (0.002)	0.69 (0.002)	0.89 (0.001)	0.41 (0.000)	0.66 (0.000)	0.67 (0.002)	0.87 (0.002)
巴拉圭	0.18 (0.000)	0.50 (0.000)	0.19 (0.002)	0.49 (0.003)	0.61 (0.003)	0.84 (0.001)	0.30 (0.000)	0.61 (0.000)	0.58 (0.002)	0.79 (0.001)

。另一个主要原因可能是表4所示的数据存在行为差异，导致模型难以在数据多样性未知的情况下学习其表示。不同国家的数据集之间存在分布偏移[6]，或许正是这一现象的原因。当各国的传感器特征分布和真实值分布（我们在第4节讨论了真实值分布）不同时，可能会产生平均效应，进而导致模型性能不如针对各个国家分别训练的模型。此外，值得注意的是，针对性能比较而训练特定国家模型和多国家模型的研究并不多 [89]。在我们找到的少数其他研究中 [52], , 与我们在复杂日常活动推断中发现的情况类似，使用特定国家模型时，基于智能手机传感器数据的人格特质推断性能更好。最后，从以人为中心的视角来看，招募参与者来收集智能手机感知数据并构建机器学习模型时，与其着眼于从单一国家招募大规模样本，不如从多个不同国家招募数量适中的参与者，这有助于将性能更好的模型部署到多个国家。

6.2 国家无关方法的泛化问题（研究问题 3）

我们分两个阶段研究了这一研究问题，具体如表1所示。在第一阶段，为评估国家特定模型泛化到新国家的程度，我们分别使用单个国家的数据训练模型，然后在其他四个国家上进行测试。在第二阶段，为评估使用四个国家训练的模型泛化到剩余国家的程度，我们采用不同的国家组合进行训练，并在剩余国家上进行测试。

• **第一阶段：** 图8总结了群体级模型的结果，图9总结了混合模型的结果。为便于比较，两幅图中提到的某个国家在同一国家进行测试时的性能结果，均来自表5。例如，根据表5，人口层级模型在意大利的 AUROC 为 0.71；图 8 中训练国家和测试国家均为意大利时标出了这一结果。人口层级的结果表明，国家无关方法在地理位置上靠近模型最初训练所在国家的国家中往往表现更好。例如，在人口层级设置下，意大利模型在意大利人群中的 AUROC 为 0.71，在丹麦（AUROC：0.69）和英国（AUROC：0.67）的表现优于在蒙古（AUROC：0.62）或巴拉圭（AUROC：0.62）的表现。混合模型也呈现出类似结果：意大利模型在丹麦和英国的表现更好。这一观察表明，来自同一地理区域（欧洲）各国的大学生，其行为可能会转化为相似的智能手机使用方式和情境，尤其是在进行类似活动期间。这与上述描述性分析中的观察结果一致：偏离总体趋势的国家通常位于欧洲以外。总而言之，即使在收集移动感知数据时采用相同的实验协议，我们仍能从模型在地理位置相距较远的国家中的性能表现观察到数据分布偏移。

• **第二阶段：** 第二阶段考察了如何扩展第一阶段的工作。我们并非在新国家测试特定国家模型，而是希望在新国家测试一个已接触过多样化数据（例如来自四个国家的数据）的模型。我们展示了随机森林模型的结果（因为它们在各项实验中的表现最佳）：训练集由四个国家的数据组成，测试集则包含第五个国家的数据。按照先前研究的建议 [52], , 各国对训练集的贡献在数据量上相等，这意味着我们必须将各国的数据下采样至相同数量（即某一国家可用数据点的最小数量）。表6展示了第二阶段实验的结果。与前几种情况类似，我们观察到性能从人口层级模型到混合模型有所提升。

[6]https://huyenchip.com/2022/02/07/data-distribution-shifts-and-monitoring.html#data-shifts

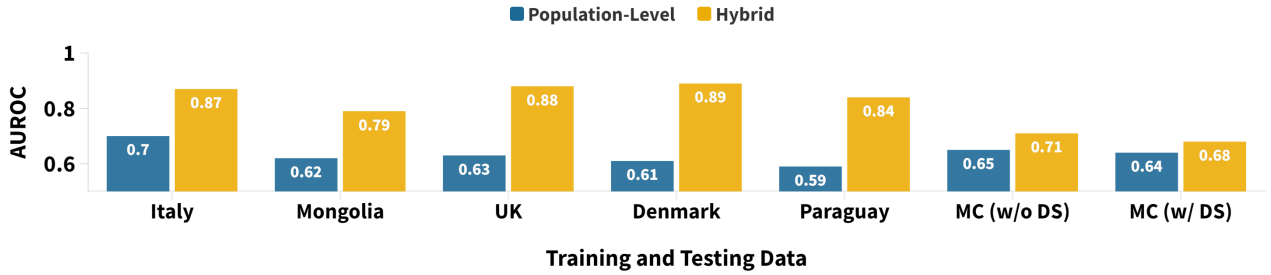


图7：AUROC 均值对比 country-specific 和多国方法在人口层级及混合模型中的比较 id 模型。MC：多国；w/o DS：不进行下采样；w/ DS：进行下采样。

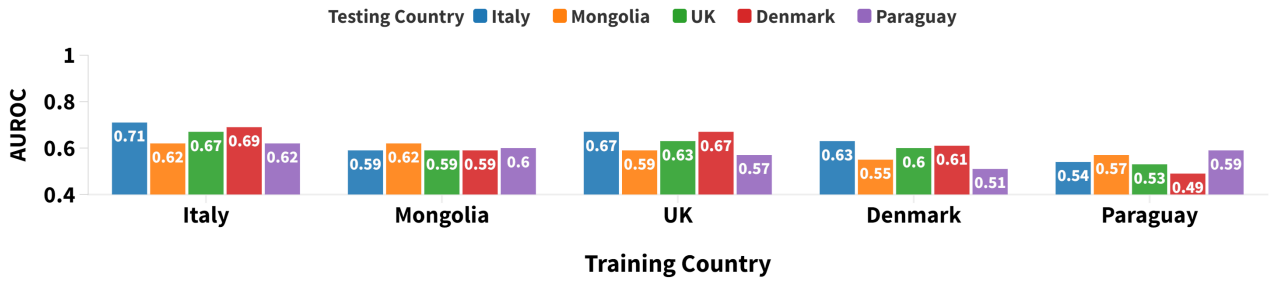


图8：在国家无关方法中使用总体层面模型获得的平均 AUROC 分数。

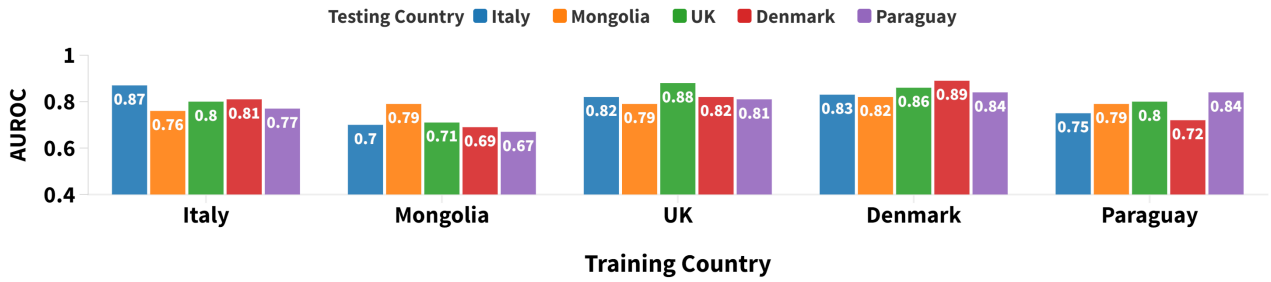


图9：在国家无关方法中使用混合模型获得的平均 AUROC 分数。

更一般地说，从 F1分数和 AUROC 分数来看，国家无关方法中的混合模型性能低于同一模型在国家特定方法中的性能。这在一定程度上可以预料，因为将来自其他分布（即其他国家）的数据纳入训练集会增加数据的方差，使模型更难准确地表示所有分布。性能下降也可能是下采样造成的。例如，在一个使用四个国家（包括意大利和巴拉圭）进行训练的模型中，与最小的巴拉圭相比，意大利在数据集中的占比最大。当把各国的样本数量减少到与巴拉圭相同的数量时，其他国家会损失大量信息：国家规模越大，原始数据集越是如此，损失就越大。这或许可以解释国家无关模型在意大利和蒙古性能较低的原因，尤其是在混合设置中。

此外，在比较不同的建模方法时，采用多国且不进行下采样所得到的结果，与国家无关方法第一阶段（混合）的结果相似，这一点是意料之中的，因为两者的训练集相似。然而，与之前的情况相比，从人口层级模型转向混合模型时，性能提升在这里并不明显。此外，多国且进行下采样的模型性能劣于前一种方法，AUROC 为 0.68，而前一种方法为 0.71。这可能是因为下采样导致我们丢失了许多国家的大量数据，从而降低了模型的表征能力。总而言之，一个混合模型

表6：通过使用来自新国家的数据测试每个国家无关模型（在四个国家进行训练）所获得的 F1分数和 AUROC 分数的均值（$\bar{S}$）和标准差（S_σ）。结果以 $\bar{S}$（S_σ）的形式呈现，其中 S 是两个指标中的任意一个。

	人口层级		混合	
测试国家	*F1*	*AUROC*	*F1*	*AUROC*
意大利	0.33 (0.005)	0.65 (0.006)	0.37 (0.004)	0.71 (0.002)
蒙古	0.30 (0.011)	0.60 (0.004)	0.37 (0.006)	0.67 (0.003)
UK	0.29 (0.004)	0.63 (0.005)	0.47 (0.004)	0.78 (0.002)
丹麦	0.38 (0.006)	0.65 (0.006)	0.63 (0.008)	0.86 (0.004)
巴拉圭	0.28 (0.005)	0.59 (0.006)	0.55 (0.006)	0.80 (0.008)

在国家无关方法中，模型无法比其国家特定对应方法更好地预测复杂活动。此外，尽管更多数据通常意味着更好的性能，但当数据遵循不同的分布时（本例中每个国家各不相同），情况并非如此。这表明每个国家都有其独特特征，使得学习统一的表示变得困难。

这些结果为我们的**研究问题3**提供了启示：在特定国家训练的复杂活动识别模型通常能够较好地泛化到其他国家（尤其是使用混合模型时）。然而，其性能仍无法与国家特定方法相媲美，这表明国家之间仍存在分布偏移。事实上，在第4节中，我们讨论了推断中使用的标签（即图 4 所示的复杂日常活动，如休息、学习、阅读等）在五个国家中的分布各不相同。此外，泛化程度往往取决于国家在地理上是否较为接近（即是否位于欧洲）。这一结果与先前研究的发现一致[52, 89]，这些研究指出了地理维度（即数据收集所在国家）对移动感知模型性能的影响。例如，[52] 发现，以移动感知数据为输入的特定国家模型能够很好地推断三种人格特质——外向性、宜人性和尽责性。此外，我们还想强调，正如近期两项研究所指出的，分布偏移和泛化问题仍是多模态移动感知领域尚未解决的问题；这两项研究考察了在同一国家不同时间段收集的类似数据集[1, 128]。这可能是因为随时间推移发生的行为变化，导致传感器数据和真实值的分布出现差异。我们的结果进一步表明，即使数据是在同一时间段内、按照相同协议收集的，国家差异仍可能导致分布偏移。

6.3 复杂日常活动识别的特征重要性

我们在实验中训练的随机森林模型会自动提供训练过程中所见特征的基尼重要性[13]。在图10中，每组箱线图表示在国家特定方法和多国方法（不含DS）下，混合模型中给定模态（如表3所定义）的特征重要性分布。初步观察发现，对于所有感知**模态**，丹麦模型中的一个蜂窝特征比其他模型中的蜂窝特征更重要。与其他国家相比，时间特征、Wi-Fi 特征和通知特征在巴拉圭更为重要。除了少数离群值外，应用事件大多不重要。考虑到用于分析的应用类型繁多，而参与者通常只频繁使用其中几种类型（例如娱乐、社交、教育、健康与健身等），这一点是合理的。我们的分析表明，这里的离群值实际上是参与者使用最多的应用。通过观察每组箱线图的上须可以发现，整体而言，预测能力最强的特征属于以下类别：时间与日期、Wi-Fi、应用使用、简单活动类型和位置。

7 讨论

7.1 结果总结

我们研究了一个多国智能手机感知数据集，以开发复杂日常活动的推断模型。我们的首要目标是探究能否利用来自智能手机的多模态传感器数据，训练出表现合理的复杂日常活动识别模型。接着，我们旨在识别通过智能手机使用所体现的国家间差异，并利用这些差异判断构建国家特定模型还是通用多国家模型是否合理，以及模型能否良好泛化。我们认为，在地理环境多样的场景中设计和部署基于感知与机器学习的应用和系统时，这些发现十分重要。三个研究问题的主要发现总结如下：

- **研究问题1**：每个国家的不同特征都可以表征一项活动。它们在一天中的分布也因国家而异，并似乎会受到影响。这一发现表明，如果在涉及人和智能手机的研究设计及数据收集阶段未采取适当的预防措施，偏差可能会不断累积。在第 7.2.1 节中，我们将围绕以下几类偏差对此进行更详细的讨论：构念偏差 [47], 样本偏差 [70], 设备类型偏差 [11], 以及用户使用习惯造成的偏差 [118]。

- **研究问题2**：采用国家特定方法训练模型，以根据智能手机数据推断 12 种复杂活动是可行的。此外，在各国内进行个性化处理可以提升性能（AUROC 范围为 0.79–0.89）。因此，国家特定方法优于多国方法；后者使用混合模型时，AUROC 仅为 0.71。不过，仅凭传感特征构建多国模型并非易事，可能需要在数据平衡和特征选择方面投入更多精力。我们的结果还表明，可以使用国家特定的部分个性化机器学习模型来捕捉疫情时代的久坐生活方式。此外，我们还表明，多模态智能手机传感器可用于识别复杂日常活动，从而实现超越二元推断的 12 类推断。在第 7.2.2 节中，我们将讨论现实生活研究为何对捕捉复杂的新兴生活方式至关重要；在第 7.2.3 节中，我们还将探讨复杂日常活动如何有助于设计新颖的情境感知应用。

- **研究问题3**：在国家无关方法下，我们发现模型能够较好地泛化到新国家。然而，正如预期的那样，其性能不及在模型训练所在的同一国家进行测试时的表现

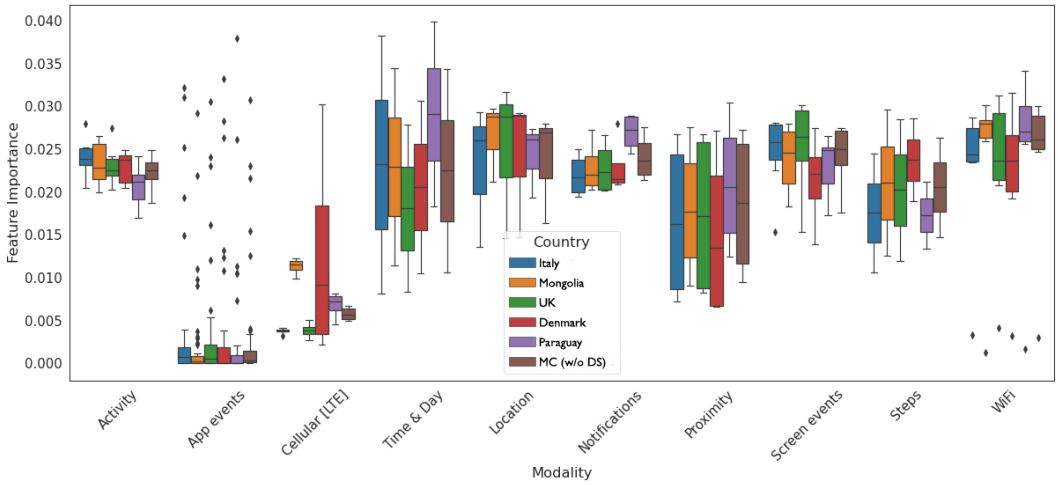


图10：特征　　混合国家特定模型和多国家模型中各特征类别的重要性　　s.

。有趣的是，在欧洲国家训练的模型，在其他欧洲国家的表现优于在巴拉圭或蒙古的表现。这个问题广泛属于领域偏移这一主题，而这一主题在移动感知文献中仍未得到充分探讨。我们将在第 7.2.4 节中对此作进一步阐述。

7.2 启示

我们的研究对理论和实践方面都有启示。

7.2.1 在研究设计中考虑国家偏差（研究问题*1*）。使用来自地理分布多样的样本（即不同国家）的感知数据开展的研究，应考虑并了解研究不同阶段可能出现的偏差来源。我们的研究以及此前有关人类行为、社会学和心理学的研究，有助于详细理解这些方面。例如，可以使用以下分类体系来描述此类偏差 [89]。（*i*）构念偏差是指目标因各国规范或环境因素而在不同国家存在不同表述 [47]。例如，在体育锻炼并不普及的国家，某种“步行”活动可能会被标记为“步行”；而在许多人将其视为健身活动的国家，它也可能被标记为一种“运动”。因此，某些行为可能是特定环境或人群所特有的。（*ii*）样本偏差涉及不同样本的可比性，而这可能受到各国招募过程的影响 [70]。例如，如果各国样本在年龄或性别方面存在差异，那么感知数据在不同国家的分布很可能也不相似。（*iii*）设备类型偏差源于不同国家参与者所使用设备的差异，以及影响传感器测量值的环境因素 [11]。世界各地的设备并未配备相同的软件和硬件，而且相似的传感器在准确度和精确度方面也可能有所不同（例如，苹果设备在发达国家更为普遍，而其他国家则普遍使用安卓手机）。最后，（*iv*）用户使用习惯导致的偏差是指不同国家的参与者使用其手机的方式不同 [118]。这类例子不胜枚举：手机的实际携带方式可能会扭曲测量结果；为节省电池或移动数据而关闭传感器（尤其是在移动数据套餐通常并非无限量的国家），也会改变测量内容；不同国家的人们使用特定应用的动机不同，也会改变最终生成的日志 [58]。Phan 等人 [89] 提出了一系列缓解策略，旨在减少偏差，推动公平的多样性感知研究。为实现这些目标，作者建议在研究的规划阶段和实施阶段采取若干措施。在规划阶段，研究人员应在本地知情人士的协助下，了解潜在的跨国差异及相关环境因素。此外，研究人员应确保研究目标在各国之间具有可比性，并且在所研究的每个国家都存在。在实施阶段，作者建议采用包容性招募策略，使每个抽样国家的样本都能代表相应的目标群体。这些建议有助于推动多样性感知研究，并缓解可能导致结果偏差的风险。

7.2.2 现实生活研究中捕捉到的活动（研究问题*2*）。从理论意义来看，值得强调的是，我们所考察的活动都属于复杂行为，通常无法在实验室研究中捕捉到。细粒度的感知式活动识别研究有助于提升对简单活动（例如步行、跑步、坐姿活动、爬楼梯等）的识别性能；这类活动在传感器数据中具有重复性，可在实验室环境中捕捉。相比之下，要构建能够捕捉复杂日常行为的基于传感的机器学习模型，就需要开展现实生活研究。学习、上课或购物等活动很难在实验室环境中复现。此外，尽管简单活动可能不会导致不同国家的模型性能出现差异，但与文化、国家层面或地理规范紧密相关的复杂日常活动会导致行为差异，进而造成感知数据的差异。在这一情境下，我们认为，先前的相关研究对这一方面的关注还不够，

在我们看来，这一领域的研究还没有充分关注这一方面。更重要的是，我们认为，研究必须从多元社区中采集数据，才能构建适用于所有目标用户的模型。尽管这是一项艰巨的任务，但若要推动该研究领域走向成熟，使其适用于更多现实应用场景，这项工作十分必要。我们的研究是朝这一方向迈出的首批尝试之一。

7.2.3 上下文感知移动应用的新型应用（研究问题*2*）。从实际意义来看，我们的研究结果表明，应将情境感知功能加入移动应用中。当前的移动应用会基于位置和简单的活动感知，提供情境感知服务、干预措施和通知 [70, 75]。然而，借助复杂日常活动识别，一系列超越当前功能范围的潜在应用将成为可能。例如，以前，智能手机只能知道用户正坐在某个特定地点。通过复杂活动识别，它可以判断用户是在学习、听课、阅读还是进食，而这些活动都需要坐着。例如，如果学生正在阅读、学习或听课，自动将电话设为静音或勿扰模式可能会很有用，尽管在许多情况下，人们会忘记这样做。总之，复杂日常活动识别未来可以提供多种使用场景，以此为基础构建移动应用。

7.2.4 多模态移动感知的领域自适应（研究问题*3*）。另一项理论意义可以从机器学习的角度来阐述。我们讨论了智能手机传感器数据集中泛化和领域偏移所面临的挑战，并说明了这种偏移如何影响模型性能，尤其是在利用多模态传感器进行复杂日常活动识别时。尽管偏差、分布偏移和模型泛化已在自然语言处理 [36], 语音[115], 和计算机视觉等其他领域得到广泛研究，智能手机感知研究却尚未受到足够关注 [63], 。我们证明，模型个性化（混合设置）可以在一定程度上减少分布偏移。从某种意义上说，按照迁移学习相关术语，这种方法类似于针对特定用户微调已训练好的模型，以实现模型个性化 [20]。以往研究也采用过此类个性化策略 [74]。然而，近期领域自适应研究指出，域适应在移动感知领域存在局限性，尤其是在处理时间序列数据时 [127]。可穿戴设备佩戴位置的多样性是人体活动识别中长期存在的问题，会影响识别模型的性能 [19, 65]。Wilson 等人 [127]研究了从不同年龄组个体采集的数据集中的域适应问题，但其研究结果仅适用于较简单的时间序列加速度计数据。其他研究也承认，目前针对智能手机和可穿戴设备数据的域适应与泛化问题缺少解决方案，这为未来探索提供了机会 [1, 128]。我们通过证实，多国多模态智能手机传感器数据需要领域自适应技术，为相关文献增添了新的发现。此外，即使在基础层面，该领域也缺少能够分别量化多模态感知特征与目标标签（如活动、情绪、社会情境等）跨数据集分布差异的方法。对这些方面的研究有助于我们更好地理解传感器中的分布偏移

数据，从而在多模态环境中更有效地运用领域自适应技术加以应对。

7.3 局限性

尽管该数据集涵盖了来自三个大洲五个不同国家的数据，但其他国家和大洲的学生行为可能与我们目前观察到的有所不同。此外，虽然我们发现，在欧洲，地理位置较近的国家表现良好，但对于其他地区也需要验证这一发现，因为地理位置较近的国家可能有着截然不同的行为和规范（例如印度和中国）。此外，在数据收集期间，不同国家的天气状况可能略有差异。本研究提及的五个国家都经历不同的季节，因为它们都离赤道有一定距离。因此，我们可以预期不同季节的特征会发生变化。然而，在实践中，很难在相似的天气状况下收集数据。

在围绕自我报告汇总传感器数据时，参与者填写自我报告期间对应的数据会被视为其当时所进行活动的一部分。如果时间窗口足够小，这种噪声可能会影响识别任务。然而，尽管当我们的目标是在细粒度感知任务中提升模型性能时，这种噪声可能会影响结果，但我们认为，就多样性意识方面的发现而言，这种噪声不会对结果产生显著影响。此外，值得注意的是，我们采用表格数据集对方法进行建模的方式，与以往在现实生活中开展的普适计算/移动感知研究类似 [70] ，因为我们没有连续真实标签。因此，这限制了我们将任务建模为时间序列问题；而大多数活动识别研究[114] 在具备连续且准确的真实测量值时，都是采用这种方式。因此，应结合本研究的探索性特征来解读结果。

此外，值得注意的是，随着不同国家人们的技术使用方式和习惯发生变化，我们可以预期模型中使用的一些信息量较大的特征也会随时间改变 [1, 128]。例如，某些国家流媒体服务使用率低的一个原因（第4节讨论过）是缺少有关使用非法下载内容的法律（例如，德国有严格的法律禁止使用非法下载内容 [97]）。各国法律的变化可能会改变年轻成年人的行为。此外，互联网价格也可能影响流媒体服务的使用情况。虽然基于带宽计费且价格低廉的互联网在发达国家很常见，但亚洲、非洲和南美洲的发展中国家并非如此；这些地区的互联网使用成本高昂，因此降低了人们使用流媒体的意愿。此外，收入水平也可能在很大程度上影响所采集的特征。例如，随着收入水平提高（这通常发生在一个国家的国内生产总值发生变化时），年轻成年人可能会更多地使用可穿戴设备进行健身追踪，从而开始在手机上使用健康与健身应用。

各国的数据量极不平衡。为公平地代表每个国家，确保每个国家的参与者人数和自我报告数量相同，可以让分类器同等地学习区分来自各国的类别。然而，目前的数据集中，意大利和蒙古占据主导地位。若操作不慎，下采样会导致

表达能力和方差的损失，使得在多国方法中难以辨别不同类别。类别标签之间也存在另一种不平衡，例如睡眠或学习等活动出现的频率高于其他活动。不过，这也合乎情理，因为我们并不期望参与者一天或一星期中的所有活动都以相同频率出现。此外，我们报告了在此类不平衡情形下更适用的 F1分数和 AUROC 分数。

最后，数据集是在2020年11月新冠疫情期间收集的，当时由于居家工作／学习的限制，大多数学生都待在家中。这解释了为什么统计分析中发现的大多数相关特征都是屏幕事件和应用事件。虽然一些相关特征与接近度和 WiFi 传感器有关，但与活动和位置有关的特征极少，除非该活动属于体育活动。这可能是因为当时的情境极力限制人们的出行。从另一个角度来看，各国大学生在这一时期的行为反映了远程工作或学习安排。随着越来越多的大学和公司采用远程工作／学习文化，我们可以预期这些做法会持续多年。因此，尽管以往许多普适计算研究都使用手机使用特征和传感特征来执行活动／行为／心理特质推断任务，但我们的研究结果表明，未来在远程学习或工作环境中，手机使用特征可能会变得更加重要；久坐行为会限制位置和惯性传感器等传感器所能提供的信息。

7.4 未来工作

该数据集收集研究的参与者群体由学生组成。因此，值得探究不同年龄组的人如何使用智能手机，以及他们的日常行为如何通过这种使用方式体现出来。除了可见的多样性之外，众所周知，深层多样性属性（人类固有且不可见）也可能影响智能手机传感器数据以及人们从事的活动 [52, 103]，例如人格（通过大五人格量表测量 [34]）、价值观（通过基本价值观调查 [45] 和人类价值观调查 [104, 105] 测量）以及智力（通过多元智能量表测量 [116]）。因此，探究这类多样性属性如何影响基于智能手机的复杂活动推断模型及其他目标变量，值得进一步研究。此外，未来工作还可以探究：排除参与者填写活动自我报告所耗时间内对应的感知数据后，分类性能会受到怎样的影响。最后，针对跨时间和国家的多模态智能手机传感器数据进行域适应，仍是未来工作中值得研究的重要问题。

8 结论

在本研究中，我们使用从学生智能手机收集的超过 21.6 万条自我报告和被动感知数据，考察了来自意大利、蒙古、英国、丹麦和巴拉圭的 637 名学生的日常行为。本研究的主要目标首先是探究能否利用多模态智能手机传感器数据推断复杂日常活动，而这将有助于开发情境感知应用；其次是考察模型能否很好地泛化到不同国家。我们有几项主要发现：*(i)* 尽管各国的活动日分布各不相同，但我们可以观察到，欧洲地理位置较近的国家之间存在相似之处。此外，一天中的时间、星期、屏幕事件和应用使用事件等特征能够指示大多数日常活动；*(ii)* 在国家特定且个性化的设置中，利用被动感知特征可以以合理的性能识别 12 种复杂日常活动。然而，与国家特定设置相比，将其扩展为多国家模型的表现并不理想；而*(iii)* 模型无法很好地泛化到其他国家（至少与国内表现相比是如此），尤其是地理位置相距较远的国家。我们需要开展更多相关研究，探讨复杂日常活动识别以及其他目标变量（例如情绪、压力、疲劳、饮食行为、饮酒行为、社会情境推断等），以验证这些发现。因此，我们认为，提升地理多样性意识对于推进移动感知和人类行为理解至关重要，有助于在多样化国家中实现更贴近现实世界的应用价值。在研究设计方面，我们主张从多样化地区和人群收集数据，以构建代表性更好的机器学习模型。在机器学习方面，我们主张开发领域适应技术，以更好地处理从不同国家收集的多模态移动感知数据。

致谢

本研究由欧盟地平线2020 WeNet项目资助，资助协议编号为823783。我们衷心感谢来自世界各地、参与本研究的所有志愿者。

参考文献

[1] Daniel A Adler、Fei Wang、David C Mohr 和 Tanzeem Choudhury。2022。用于被动心理健康症状预测的机器学习：跨不同纵向移动感知研究的泛化。《PLOS ONE》17，第4期（2022），e0266516。[2] Dhoha A Alsaleh、Michael T Elliott、Frank Q Fu 和 Ramendra Thakur。2019。社交媒体采纳中的跨文化差异。《互动营销研究杂志》（2019）。[3] Tim Althoff、Rok Sosič、Jennifer L Hicks、Abby C King、Scott L Delp 和 Jure Leskovec。2017。大规模身体活动数据揭示全球活动不平等。《自然》547，第7663期（2017），336–339。[4] Carlos Avilés-Cruz、Andrés Ferreyra-Ramírez、Arturo Zúñiga-López 和 Juan Villegas-Cortéz。2019。用于人体活动识别的粗细卷积深度学习策略。《传感器》19，第7期（2019年3月），1556。https://doi.org/ 10.3390/s19071556[5] Sangwon Bae、Denzil Ferreira、Brian Suffoletto、Juan C Puyana、Ryan Kurtz、Tammy Chung 和 Anind K Dey。2017。利用基于智能手机的传感器检测年轻成年人的饮酒事件。ACM交互式、移动、可穿戴与泛在技术论文集1，第2期（2017），1–36。[6] Sangwon Bae、Denzil Ferreira、Brian Suffoletto、Juan C. Puyana、Ryan Kurtz、Tammy Chung 和 Anind K. Dey。2017。利用基于智能手机的传感器检测年轻成年人的饮酒事件。ACM交互式、移动、可穿戴与泛在技术论文集1，第2期（2017年6月），1–36。https: //doi.org/ 10 1145 3090051 。/ Ling Bao 和 Stephen S. Intille。[7] 2004。基于用户标注加速度数据的活动识别。载于普适计算，Alois Ferscha 和 Friedemann Mat- tern（编）。Springer Berlin Heidelberg，德国柏林、海德堡，1–17。[8] Naomi S. Baron 和 Ylva Hård af Segerstad。2010。手机使用中的跨文化模式：瑞典、美国和日本的公共空间与可联系性。《新媒体与社会》12，第1期（2010），13–34。https://doi.org/10.1177/ 1461444809355111 10 1177 1461444809355111 [9] arXiv:https://doi.org/ 。/ Joan-Isaac Biel、Nathalie Martin、David Labbe 和 Daniel Gatica-Perez。2018。Bites'n'Bits：从情境移动数据推断饮食行为。ACM交互式、移动、可穿戴与泛在技术论文集1，第4期（1月 2018 1 33 10 1145 3161161），– 。https://doi.org/ 。/ Ulf [10] Blanke 和 Bernt Schiele。 。2009 。通过活动检测进行日常惯例识别。载于位置与情境感知，Tanzeem Choudhury、Aaron

Quigley、Thomas Strang 和 Koji Suginuma（编）。第5561卷。Springer Berlin Heidelberg，德国柏林、海德堡，192–206。https://doi.org/10.1007/978-3-642- 01721-6_12 系列标题：《计算机科学讲义》。[11] Henrik Blunck、Niels Olof Bouvin、Tobias Franke、Kaj Grønbæk、Mikkel B Kjaer- gaard、Paul Lukowicz 和 Markus Wüstenberg。2013。面向代表性数据采集的移动感知应用中的异质性。载于2013年ACM普适与泛在计算会议附属出版物论文集。1087–1098。[12] Emma Bouton-Bessac、Lakmal Meegahapola 和 Daniel Gatica-Perez。2022。口袋中的一天：基于智能手机加速度计的复杂活动识别。第16届EAI国际医疗保健普适计算技术会议（PervasiveHealth）论文集（2022）。[13] Leo Breiman（编）。1998。分类回归树（第1版，crc press 重印版）。Chapman & Hall/CRC，美国佛罗里达州博卡拉顿。[14] Leo Breiman。2001。随机森林。机器学习45，第1期（2001），5–32。https://doi.org/ 10 1023 1010933404324 。/A：Naomi Canton。 。手机文化：文化[15] 2012 差异如何影响移动设备使用。https://edition.cnn.com/2012/09/27/tech/mobile-culture-usage/index.html [16] Luca Canzian 和 Mirco Musolesi。2015。抑郁轨迹：通过分析智能手机移动轨迹对抑郁状态进行非侵入式监测。载于2015年ACM普适与泛在计算国际联合会议论文集。1293–1304。[17] 皮尤研究中心。2021。移动设备概况。https://www.pewresearch.org/ internet/fact-sheet/mobile/[18] 皮尤研究中心。2017。通过了解用户正在做什么来调整您的应用程序。2022年12月12日检索自 https://www.pewresearch.org/fact-tank/2017/09/13/about-6-in-10-young-adults-in-u-s-primarily-use-online-streaming-to-watch-tv/[19] Youngjae Chang、Akhil Mathur、Anton Isopoussu、Junehwa Song 和 Fahim Kawsar。2020。面向稳健人类活动识别的无监督域适应系统研究。ACM交互式、移动、可穿戴与泛在技术论文集4，第1期（2020），1–30。[20] Yiqiang Chen、Xin Qin、Jindong Wang、Chaohui Yu 和 Wen Gao。2020。Fed- health：一种用于可穿戴医疗保健的联邦迁移学习框架。IEEE智能系统35，第4期（2020），83–93。[21] Matty Chiva。1997。饮食的文化层面与进餐频率。英国营养学杂志77，S1（1997），S21–S28。[22] François Chollet 等。2015。Keras。https://keras.io。[23] Tanzeem Choudhury、Gaetano Borriello、Sunny Consolvo、Dirk Haehnel、Beverly Harrison、Bruce Hemingway、Jeffrey Hightower、Predrag "Pedja" Klasnja、Karl Koscher、Anthony LaMarca、James A. Landay、Louis LeGrand、Jonathan Lester、Ali Rahimi、Adam Rea 和 Danny Wyatt。2008。移动感知平台：一个嵌入式活动识别系统。IEEE普适计算7，第2期（4月 2008 32 41 10 1109 2008 39），– 。https://doi.org/[24] 。/MPRV。 。Daniel G Cobian、Nicole S Daehn、Paul A Anderson 和 Bryan C Heiderscheit。2013。健康年轻成年人在进行日常生活活动期间颈椎和腰椎的主动活动范围。Spine 38，第20期（2013），1754–1763。[25] Lynn Coorevits 和 Tanguy Coenen。2016。可穿戴健身追踪器的兴衰。管理学会论文集2016，第1期（2016年1月），17305。https: //doi.org/10.5465/ambpp.2016.17305abstract [26] Victor P Cornet 和 Richard J Holden。2018。基于智能手机的健康与福祉被动感知系统综述。生物医学信息学杂志77（2018），120–132。[27] AI Cultures。2022。NeurIPS 2022 AI Cultures 研讨会。2023年2月12日检索自 https://ai-cultures.github.io/ [28] Anthony Cuthbertson。2019。研究称，自动驾驶汽车更有可能撞向黑人。https://www.independent.co.uk/life-style/gadgets-and-tech/news/self-driving-car-crash-racial-bias-black-people-study-a8810031.html[29] Vedant Das Swain、Victor Chen、Shrija Mishra、Stephen M Mattingly、Gregory D Abowd 和 Munmun De Choudhury。2022。通过被动感知预测心理福祉中的语义鸿沟。载于计算机系统中的人为因素会议。1–16。[30] Yves-Alexandre de Montjoye、Laura Radaelli、Vivek Kumar Singh 和 Alex " Sandy" Pentland。2015。购物中心中的独特性：论信用卡元数据的重新识别性。《科学》347，第6221期 2015 536 539 10 1126 1256297（ ），– 。https://doi.org/ 。/ science。arXiv:https://www.science.org/doi/pdf/10.1126/science.1256297[31] Michael B Del Rosario、Kejia Wang、Jingjing Wang、Ying Liu、Matthew Brodie、Kim Delbaere、Nigel H Lovell、Stephen R Lord 和 Stephen J Redmond。2014。使用智能手机比较年轻和年长群体的活动分类。《生理测量》35，第11期（2014年11月），2269–2286。https: //doi.org/ 10 1088 0967 3334 35 11 2269 。/ - / / / Chengbin Deng、Weiying Lin、[32] Xinyue Ye、Zhenlong Li、Ziang Zhang 和 Gang- gang Xu。2018。作为每小时细粒度电力消耗估计代理的社交媒体数据。环境与规划A：经济与空间50，第8期（2018），1553–1557。

[33] Stefan Dernbach、Barnan Das、Narayanan C. Krishnan、Brian L. Thomas 和 Diane J. Cook。2012。通过智能手机识别简单和复杂活动。载于《2012年第八届智能环境国际会议》。IEEE，墨西哥瓜纳华托， 214 221 10 1109 2012 39 – 。[34] https://doi.org/ 。/IE. 。M Brent Donnellan、Frederick L Oswald、Brendan M Baird 和 Richard E Lucas。2006。迷你 IPIP 量表：简短却有效的人格大五因素测量工具。《心理评估》18，2（2006），192。[35] Nathan Eagle 和 Alex (Sandy) Pentland。2006。Reality Mining：感知复杂社会系统。《个人与泛在计算》10，4（2006年5月），255–268。https://doi.org/ 10 1007 00779 005 0046 3 。/s - - - Hady Elsahar 和 Matthias [36] 2019 Gallé。 。标注与否？预测领域偏移下的性能下降。载于《2019年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集》。2163–2173。[37] Charlie Giattino Esteban Ortiz-Ospina 和 Max Roser。2020。时间利用。Our World in Data（2020）。[38] Jonny Farringdon、Andrewj Moore、Nancy Tilbury、James Church 和 Pieter D. Biemond。1999。用于情境感知的可穿戴传感器徽章与传感器夹克。[39] Kimberly Fisher 和 John Robinson。2010。22个国家的日常生活规律：日记证据显示每日在三十种活动上花费的平均时间。技术报告。牛津大学时间利用研究中心。https://www.timeuse.org/sites/default/files/public/ctur_technical_ paper/869/CTUR_Technical_Paper_2010-01.pdf [40] Fausto Giunchiglia、Enrico Bignotti 和 Mattia Zeni。2017。个人情境建模与标注。载于《2017年IEEE普适计算与通信国际会议研讨会（PerCom研讨会）》。IEEE，117–122。[41] Fausto Giunchiglia、Ivano Bison、Matteo Busso、Ronald Chenu-Abente、Marcelo Rodas、Mattia Zeni、Can Gunel、Giuseppe Veltri、Amalia De Götzen、Peter Kun、Amarsanaa Ganbold、Altangerel Chagnaa、George Gaskell、Sally Stares、Miriam Bidoglia、Luca Cernuzzi、Alethia Hume、Jose Luis Zarza、Hao Xu、Donglei Song、Shyam Diwakar、Chaitanya Nutakki、Salvador Ruiz Correa、Andrea-Rebeca Mendoza、Lakmal Meegahapola 和 Daniel Gatica-Perez。2022。关于日常生活规律与社会实践的全球多样性试点研究（2020–2021）。特伦托大学技术报告——DataScientia数据集描述。https://iris.unitn.it/handle/11572/338382。[42] Taesik Gong、Yewon Kim、Adiba Orzikulova、Yunxin Liu、Sung Ju Hwang、Jinwoo Shin 和 Sung-Ju Lee。2021。DAPPER：移动感知中域适应的性能估计。arXiv预印本 arXiv:2111.11053（2021）。[43] Zhiqiang Gong、Ping Zhong 和 Weidong Hu。2019。机器学习中的多样性。《IEEE Access》7（2019），64323–64350。https://doi.org/10.1109/ACCESS. 2019.2917620[44] 谷歌。2022。通过了解用户正在进行的操作来调整应用。检索于2022年2月12日，网址：https://developers.google.com/location-context/activity- recognition [45] Valdiney V Gouveia、Taciano L Milfont 和 Valeschka M Guerra。2014。人类价值观的功能理论：检验其内容与结构假设。《人格与个体差异》60（2014），41–47。[46] Trevor Hastie、Saharon Rosset、Ji Zhu 和 Hui Zou。2009。多类 AdaBoost。《统计学及其交叉领域》2，3（2009），349–360。https://doi.org/10.4310/SII.2009. v2.n3.a8[47] Jia He 和 Fons van de Vijver。2012。跨文化研究中的偏差与等值性。《心理学与文化在线读物》2，2（2012），2307–0919。[48] Ida Marie Henriksen、Marianne Skaar 和 Aksel Tjora。2020。公共场所使用智能手机的构成性实践。《社会》10，4（2020）。https://doi.org/10.3390/soc10040078[49] Seyed Amir Hoseini-Tabatabaei、Alexander Gluhak 和 Rahim Tafazolli。2013。机会式用户情境识别的智能手机系统综述。《ACM计算调查》（CSUR）45，3（2013），1–51。[50] Tâm Huynh、Mario Fritz 和 Bernt Schiele。2008。使用主题模型发现活动模式。载于《第十届泛在计算国际会议论文集——UbiComp '08》。ACM出版社，韩国首尔，10。https://doi.org/ 10 1145 1409635 1409638 。/ 。Kimmo Karkkainen 和 Jungseock [51] Joo。2021。Fairface：用于偏差测量与缓解的面部属性数据集，旨在平衡种族、性别和年龄。载于《IEEE/CVF冬季计算机视觉应用会议论文集》。1548–1558。[52] Mohammed Khwaja、Sumer S. Vaid、Sara Zannone、Gabriella M. Harari、A. Aldo Faisal 和 Aleksandar Matic。2019。建模人格与建模个人特质：对五个国家真实环境中的移动数据进行分析，揭示文化对人格模型的影响。《ACM交互式、移动、可穿戴与泛在技术论文集》3，3（2019年9月），1–24。https: //doi.org/ 10 1145 3351246 。/ Hae-Young Kim。 。方差分析（ANOVA）：比较两[53] 2014 个以上组的均值。《修复牙科与牙髓学》39，1（2014），74–77。[54] Gierad Laput 和 Chris Harrison。2019。使用智能手表感知细粒度手部活动。载于《2019年计算机系统中的人为因素会议论文集》。ACM，英国苏格兰格拉斯哥，1–13。https://doi.org/10. 1145/3290605.3300568

[55] Gierad Laput 和 Chris Harrison。2019。使用智能手表感知细粒度手部活动。载于《*2019*年计算机系统中的人为因素会议论文集》。1–13。[56] Tong Li、Mingyang Zhang、Yong Li、Eemil Lagerspetz、Sasu Tarkoma 和 Pan Hui。2021。新冠肺炎对智能手机使用的影响。《*IEEE*物联网期刊》8，23（2021年12月），16723–16733。https://doi.org/10.1109/JIOT. 2021.3073864[57] Robert LiKamWa、Yunxin Liu、Nicholas D. Lane 和 Lin Zhong。2013。MoodScope：基于智能手机使用模式构建情绪传感器。载于《第*11*届年度移动系统、应用与服务国际会议论文集——*MobiSys '13*》。ACM出版社，中国台湾台北，389。https: // doi.org/10.1145/2462456.2464449[58] Soo Ling Lim、Peter J Bentley、Natalie Kanakam、Fuyuki Ishikawa 和 Shinichi Honiden。2014。研究移动应用用户行为中的国家差异及其对软件工程的挑战。《*IEEE*软件工程汇刊》41，1（2014），40–64。[59] Rex Liu、Albara Ah Ramli、Huanle Zhang、Erik Henricson 和 Xin Liu。2021。可穿戴传感器人体活动识别概述：医疗保健与人工智能。*arXiv:2103.15990 [cs, eess]*（2021年8月）。http://arxiv. org/abs/2103.15990 arXiv: 2103.15990。[60] Tiffany Liu、Javier Hernandez、Mar Gonzalez-Franco、Antonella Maselli、Melanie Kneisel、Adam Glass、Jarnail Chudge 和 Amos Miller。2022。表征并预测盲人和低视力人士对音频导航应用的参与度。载于《*2022*年计算机系统中的人为因素会议扩展摘要》（美国路易斯安那州新奥尔良）（*CHI EA '22*）。计算机协会，美国纽约州纽约市，文章411，7页。https://doi.org/ 10 1145 3491101 3519862 。/ 。Olatz Lopez-Fernandez、Daria J. [61] Kuss、Lucia Romo、Yannick Morvan、Laurence Kern、Pierluigi Graziani、Amélie Rousseau、Hans-Jürgen Rumpf、Anja Bischof、Ann-Kathrin Gässler、Adriano Schimmenti、Alessia Passanisi、Niko Männikkö、Maria Käériänen、Zsolt Demetrovics、Orsolya Király、Mariano Chóliz、Juan José Zacarés、Emilia Serra、Mark D. Griffiths、Halley M. Pontes、Bernadeta Lelonek-Kuleta、Joanna Chwaszcz、Daniele Zullino、Lucien Rochat、Sophia Achab 和 Joël Billieux。2017。年轻成年人自我报告的手机依赖：一项欧洲跨文化实证调查。《行为成瘾杂志》6，2（2017），168–177。https://doi.org/ 。020 。 [62] 2021 Amanda D Lotz。。介于全球与本地之间：描绘 Netflix 作为跨国服务的地理分布。《国际文化研究杂志》24，2（2021），195–215。[63] Yawei Luo、Liang Zheng、Tao Guan、Junqing Yu 和 Yi Yang。2019。深入考察领域偏移：用于语义一致域适应的类别级对抗样本。载于《*IEEE/CVF*计算机视觉与模式识别会议论文集》。2507–2516。[64] Stephen Makonin、Lyn Bartram 和 Fred Popowich。2013。更智能的智能家居：环境智能案例研究。《*IEEE*普适计算》12，1（2013），58–66。https://doi.org/10.1109/MPRV.2012.58[65] Akhil Mathur、Anton Isopoussu、Nadia Berthouze、Nicholas D Lane 和 Fahim Kawsar。2019。用于稳健感知系统的无监督域适应。载于《*2019*年*ACM*普适与泛在计算国际联合会议暨*2019*年*ACM*可穿戴计算国际研讨会附属论文集》。505–509。[66] Akhil Mathur、Lakshmi Manasa Kalanadhabhatta、Rahul Majethia 和 Fahim Kawsar。2017。超越市场调研：揭开印度智能手机用户行为的面纱。《*ACM*交互式、移动、可穿戴与泛在技术论文集》 1, 3 2017 1 27 10 1145 3130947 （9月），– [67] 2013 。https://doi.org/ 。/ F Joseph McClernon 和 Romit Roy Choudhury。。我就是你的智能手机，我知道你即将吸烟：移动感知与计算方法在吸烟研究和治疗中的应用。《尼古丁与烟草研究》15，10（2013），1651–1654。[68] Lakmal Meegahapola、Wageesha Bangamuarachchi、Anju Chamantha、Salvador Ruiz-Correa、Indika Perera 和 Daniel Gatica-Perez。2022。在情境中感知进食事件：仅使用智能手机的方法。《*IEEE Access*》10，文章（2022）。[69] Lakmal Meegahapola、William Droz、Peter Kun、Amalia de Götzen、Chaitanya Nutakki、Shyam Diwakar、Salvador Ruiz Correa、Donglei Song、Hao Xu、Miriam Bidoglia、George Gaskell、Altangerel Chagnaa、Amarsanaa Ganbold、Tsolmon Zundui、Carlo Caprini、Daniele Miorandi、Alethia Hume、Jose Luis Zarza、Luca Cernuzzi、Ivano Bison、Marcelo Rodas Britez、Matteo Busso、Ronald Chenu-Abente、Can Günel、Fausto Giunchiglia、Laura Schelenz 和 Daniel Gatica-Perez。2023。基于移动感知的情绪推断模型的泛化与个性化：对八个国家大学生的分析。《*ACM*交互式、移动、可穿戴与泛在技术论文集》6，4，文章176（2023年1月），32页。[70] Lakmal Meegahapola 和 Daniel Gatica-Perez。2020。面向年轻成年人身心健康的智能手机感知：综述。《*IEEE Access*》9（2020），3374–3399。[71] Lakmal Meegahapola、Florian Labhart、Thanh-Trung Phan 和 Daniel Gatica-Perez。2021。考察年轻成年人饮酒的社会情境

智能手机感知与成年人。*ACM*交互式、移动、可穿戴与泛在技术论文集5, 3（2021年9月），1–26。https: //doi.org/ 10 1145 3478126 . / Lakmal [72] Meegahapola、Salvador Ruiz-Correa 和 Daniel Gatica-Perez。2020。独自一人还是与他人同行？通过移动感知了解大学生的进食事件。载于第*19*届移动与普适多媒体国际会议。ACM，德国埃森，162–166。https://doi.org/10.1145/3428361. 3428463 [73] Lakmal Meegahapola、Salvador Ruiz-Correa 和 Daniel Gatica-Perez。2020。保护移动饮食日记，避免记录过多个人信息。载于第*19*届移动与普适多媒体国际会议。ACM，德国埃森，212–222。https://doi.org/ 10 1145 3428361 3428468 . / 。Lakmal Meegahapola、Salvador [74] Ruiz-Correa、Viridiana del Carmen Robledo-Valero、Emilio Ernesto Hernandez-Huerfano、Leonardo Alvarez-Rivera、Ronald Chenu-Abente 和 Daniel Gatica-Perez。2021。再吃一口？利用智能手机感知和自我报告推断大学生的食物摄入量。*ACM*交互式、移动、可穿戴与泛在技术论文集 5, 1 2021 1 28 10 1145 3448120 （3月），–。 [75] 2017 https://doi.org/。/ Abhinav Mehrotra 和 Mirco Musolesi。智能通知系统：现状综述与研究挑战。*arXiv*预印本 *arXiv:1711.10171*（2017）。[76] Kathryn Mercer、Lora Giangregorio、Eric Schneider、Parmit Chilana、Melissa Li 和 Kelly Grindrod。2016。50岁以上且患有慢性疾病的成年人对市售可穿戴活动追踪器的接受度：混合方法评估。*JMIR mHealth and uHealth* 4, 1（2016年1月），e7。https://doi.org/ 10 2196 4225 . /mhealth。Mike A Merrill 和 [77] 2022 Tim Althoff。自监督预训练与迁移学习助力基于小型移动感知数据集预测流感和新冠肺炎。*arXiv*预印本 *arXiv:2205.13607*（2022）。[78] Daniel Miller、Laila Abed Rabho、Patrick Awondo、Maya de Vries、Marília Duque、Pauline Garvey、Laura Haapio-Kirk、Charlotte Hawkins、Alfonso Otaegui、Shireen Walton 和 Xinyuan Wang。2021。全球智能手机：超越青年科技。UCL Press。http://www.jstor.org/stable/j.ctv 1 0 1 b fvh Chan [79] 2020 Naseeb 和 Bilal Al Saeedi。基于深度学习的运动与交通数据集活动识别。载于*2020*年*ACM*普适与泛在计算国际联合会议暨*2020*年*ACM*可穿戴计算国际研讨会附属论文集。ACM，虚拟活动，墨西哥，329–334。https://doi.org/10.1145/3410530.3414348[80] Tobias Nef、Prabitha Urwyler、Marcel Büchler、Ioannis Tarnanas、Reto Stucki、Dario Cazzoli、René Müri 和 Urs Mosimann。2015。评估三种最先进的分类器，以利用智能家居环境数据识别日常生活活动。《传感器》15, 5（2015年5月），11725–11740。https://doi.org/ 10.3390/s150511725[81] Justin J. Nelson 和 Christopher M. Pieper。2022。"无限抱负的病症"：智能手机、意义寻求与失范生成。《社会学视角》（2022年8月），073112142211142。https://doi.org/10.1177/07311214221114296[82] Subigya Nepal、Shayan Mirjafari、Gonzalo J Martinez、Pino Audia、Aaron Striegel 和 Andrew T Campbell。2020。利用移动感知检测信息工作者的职位晋升。*ACM*交互式、移动、可穿戴与泛在技术论文集 4, 3 2020 1 28 10 1145 3414118 （），–。 [83] 2020 https://doi.org/。/ Johnny Nhan、Kendra Bowen 和 Aaron Bartula。比较一所公立大学与一所私立大学中低价流媒体服务和收入对电影盗版的影响。《技术与社会》60（2020），101213。[84] Mikio Obuchi、Jeremy F Huckins、Weichen Wang、Alex daSilva、Courtney Rogers、Eilis Murphy、Elin Hedlund、Paul Holtzheimer、Shayan Mirjafari 和 Andrew Campbell。2020。利用移动感知预测脑功能连接。*ACM*交互式、移动、可穿戴与泛在技术论文集 4, 1 2020 1 22 10 1145 3381001 （），–。 [85] 2015 https://doi.org/。/ Hyungik Oh、Laleh Jalali 和 Ramesh Jain。利用实时个人活动监测中的情境构建智能通知系统。载于*2015*年*IEEE*多媒体与博览会国际会议（*ICME*）。IEEE，1–6。[86] Jay A. Olson、Dasha A. Sandra、Élissa S. Colucci、Alain Al Bikaii、Denis Chmoule-vitch、Johnny Nahas、Amir Raz 和 Samuel P.L. Veissière。2022。全球范围内智能手机成瘾日益严重：对24个国家的荟萃分析。《人类行为中的计算机》129（2022），107138。https://doi.org/10.1016/j.chb.2021. 107138[87] S. Park、C. Gopalsamy、R. Rajamanickam 和 S. Jayaraman。1999。可穿戴主板：用于医疗应用的柔性信息基础设施或传感内衬。《健康技术与信息学研究》62（1999），252–258。[88] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duchesnay。2011。Scikit-learn：Python 机器学习。《机器学习研究期刊》 12（2011），2825–2830。[89] Le Vy Phan、Nick Modersitzki、Kim Karen Gloystein 和 Sandrine Müller。2022。全球移动感知：跨文化研究的考量。预印本。PsyArXiv。https://doi.org/10.31234/osf.io/q8c7y

[90] Thanh-Trung Phan、Florian Labhart、Skanda Muralidhar 和 Daniel Gatica-Perez。2020。通过智能手机感知和主动人类参与了解夜间大量饮酒。载于第*14*届*EAI*普适计算医疗保健技术国际会议论文集。211–222。[91] Vinodkumar Prabhakaran、Rida Qadri 和 Ben Hutchinson。2022。人工智能中的文化不一致性。*arXiv*预印本 *arXiv:2211.13069*（2022）。[92] Angshu Rai、Zhixian Yan、Dipanjan Chakraborty、Tri Kurniawan Wijaya 和 Karl Aberer。2012。通过单个智能手机加速度计，在自然环境中挖掘复杂活动。载于第六届传感器数据知识发现国际研讨会论文集。43–51。[93] Inioluwa Deborah Raji 和 Joy Buolamwini。2019。可执行审计：调查公开点名有偏差的商业人工智能产品性能结果所产生的影响。载于*2019*年*AAAI/ACM*人工智能、伦理与社会会议论文集。ACM，美国夏威夷州檀香山，429–435。https://doi.org/10.1145/3306618. 3314244[94] Haroon Rashid、Sanjana Mendu、Katharine E Daniel、Miranda L Beltzer、Bethany A Teachman、Mehdi Boukhechba 和 Laura E Barnes。2020。根据稀疏采集的移动传感器数据预测社交焦虑的主观测量值。*ACM*交互式、移动、可穿戴与泛在技术论文集4, 3（2020），1–24。[95] Zubair Ahmed Ratan、Sojib Bin Zaman、Sheikh Mohammed S hariful Islam 和 Hassan Hosseinzadeh。2021。智能手机过度使用：新冠肺炎疫情中的隐性危机。《卫生政策与技术》10, 1（2021年3月），21–22。https://doi.org/10.1016/ j.hlpt. 2021 01 002 。Ashwini S Rathod、[96] Abhishek Ingole、Abhay Gaidhane 和 Sonali G Choudhari。2022。青少年和年轻成年人过度使用智能手机相关的心理疾病：综述。*Cureus*（2022年10月）。https://doi.org/ 10 7759 30756 . /cureus。Simon Rump。你是哪种窃贼？将观念、人格特质和音乐品味与非法下载联系起来——偏好、特质和观念如何影响网络犯罪。理学学士论文。特温特大学。[97] [98] Heba Saadeh、Reem Q. Al Fayez、Assem Al Refaei、Nour Shewaikani、Hamzah Khawaldah、Sobuh Abu-Shanab 和 Maysa Al-Hussaini。2021。新冠肺炎隔离期间大学生使用智能手机：一个伦理触发因素。《公共卫生前沿》9（2021年7月），600134。https://doi.org/10.3389/fpubh. 2021.600134[99] Saguna Saguna、Arkady Zaslavsky 和 Dipanjan Chakraborty。2013。运用情境驱动活动理论和活动特征进行复杂活动识别。《*ACM*计算机与人机交互汇刊》（*TOCHI*）20, 6（2013），1–34。[100] Darshan Santani、Trinh-Minh-Tri Do、Florian Labhart、Sara Landolt、Emmanuel Kuntsche 和 Daniel Gatica-Perez。2018。DrinkSense：利用智能手机刻画青少年饮酒行为。*IEEE*移动计算汇刊 17, 10 2018 2279 2292 10 1109 2018 2797901（10月），–。[101] https://doi.org/。/TMC。Jeffer Eidi Sasaki、Amanda M. Hickey、John W. Staudenmayer、Dinesh John、Jane A. Kent 和 Patty S. Freedson。2016。自由生活老年人活动分类算法的性能。《医学与科学：体育与运动》 48, 5 2016 941 950 10 1249 0000000000000844（5月），–。[102] https://doi.org/。/MSS。Laura Schelenz、Ivano Bison、Matteo Busso、Amalia de Götzen、Daniel Gatica-Perez、Fausto Giunchiglia、Lakmal Meegahapola 和 Salvador Ruiz-Correa。2021。设计关注多样性的社会关系平台：理论、实践与伦理挑战。载于*2021*年*AAAI/ACM*人工智能、伦理与社会会议论文集（美国虚拟活动）（*AIES '21*）。计算机协会，美国纽约州纽约，905–915。https://doi.org/10. 1145/3461702.3462595[103] Laura Schelenz、Ivano Bison、Matteo Busso、Amalia de Götzen、Daniel Gatica-Perez、Fausto Giunchiglia、Lakmal Meegahapola 和 Salvador Ruiz-Correa。2021。设计关注多样性的社会关系平台：理论、实践与伦理挑战。载于*2021*年*AAAI/ACM*人工智能、伦理与社会会议论文集。ACM，美国虚拟活动，905–915。https: //doi.org/ 10 1145 3461702 3462595 . / 。Shalom H Schwartz。人类价值观的结构和[104] 内容是否存在普遍特征？《社会问题杂志》50, 4（1994），19–45。[105] Shalom H Schwartz、Gila Melech、Arielle Lehmann、Steven Burgess、Mari Harris 和 Vicki Owens。2001。采用不同测量方法拓展基本人类价值观理论的跨文化效度。《跨文化心理学杂志》32, 5（2001），519–542。[106] Scikit-Learn。2022年2月14日，2022。*Scikit-Learn*指标——*AUROC*。检索于2月，网址：https://scikit-learn.org/stable/modules/generated/sklearn.metrics. roc auc score.html[107] Scikit-Learn。2022。*Scikit-Learn*指标——*F1*分数。检索于2022年2月14日，网址：https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score. html [108] Sandra Servia-Rodríguez、Kiran K. Rachuri、Cecilia Mascolo、Peter J. Rentfrow、Neal Lathia 和 Gillian M. Sandstrom。2017。服务于心理健康的移动感知：一项大规模纵向研究。载于第*26*届万维网国际会议论文集。万维网国际会议指导委员会，澳大利亚珀斯，103–112。https://doi.org/10. 1145/3038912.3052618 [109] Craig JR Sewall、Todd M Bear、John Merranko 和 Daniel Rosen。2020年。心理社会福祉和使用时长如何预测对数字技术使用情况的回顾性估计误差。《移动媒体与传播》8，3（2020），379–399。[110] Pekka Siirtola、Perttu Laurinen、Juha Roning 和 Hannu Kinnunen。2011年。高效的基于加速度计的游泳锻炼追踪。载于*2011*年*IEEE*计算智能与数据挖掘研讨会（*CIDM*）。IEEE，法国巴黎，156 161 10 1109 2011 5949430 – 。https://doi.org/ 。/CID [111] M. 。Dale Southerton、Cecilla Díaz-Méndez、Alan Warde 等。2012。行为变化与饮食实践的时间顺序：英国—西班牙比较。《国际农业与食品社会学杂志》19，1（2012），19–36。[112] Stephanie Stockwell、Mike Trott、Mark Tully、Jae Shin、Yvonne Barnett、Laurie Butler、Daragh McDermott、Felipe Schuch 和 Lee Smith。2021年。新冠肺炎疫情封控前后身体活动和久坐行为的变化：系统综述。《BMJ开放运动与运动医学》 7, 1 2021 000960 10 1136 2020 000960（1月），e 。[113] https://doi.org/ 。/bmjsem- - Marcin Straczkiewicz、Peter James 和 Jukka-Pekka Onnela。 。面向健康研究的基于智能手机的人类活动识别方法系统综述。《npj数字医学》4，1（2021年12月），148。https://doi.org/10.1038/s41746-021-00514-4[114] Marcin Straczkiewicz、Peter James 和 Jukka-Pekka Onnela。2021年。面向健康研究的基于智能手机的人类活动识别方法系统综述。《NPJ数字医学》4，1（2021），1–15。[115] Sining Sun、Binbin Zhang、Lei Xie 和 Yanning Zhang。2017年。一种用于鲁棒语音识别的无监督深度域适应方法。《Neurocomput- ing》257（2017），79–87。[116] Kirsi Tirri 和 Petri Nokelainen。2008年。使用《多元智能剖面问卷III》识别多元智能。《心理学科学》50，2 2008 206( )，。[117] Taylor Tobin。2018年。世界上12个不同地方的人们通常在什么时间吃晚餐。https://www.insider.com/dinner-times-around-the-world-2018-11[118] Fons Van de Vijver 和 Norbert K Tanzer。2004年。跨文化评估中的偏差与等值性：概述。《欧洲应用心理学评论》54，2（2004），119–135。[119] Alexander J.A.M. van Deursen、Colin L. Bolle、Sabrina M. Hegner 和 Piet A.M. Kommers。2015年。习惯性和成瘾性智能手机行为建模：智能手机使用类型、情绪智力、社会压力、自我调节、年龄和性别的作用。《人类行为中的计算机》45（2015），411–420。https: // doi.org/ 10 1016 2014 12 039 。 /j.chb. 。 。Vincent T. van Hees、[120] Rajna Golubic、Ulf Ekelund 和 Søren Brage。2013年。研究设计对活动类型分类器开发与评估的影响。《应用生理学杂志》114，8（2013年4月），1042–1051。https://doi.org/10. 1152 00984 2012 /japplphysiol. 。Rui Wang、[121] Weichen Wang、Min SH Aung、Dror Ben-Zeev、Rachel Brian、An- drew T Campbell、Tanzeem Choudhury、Marta Hauser、John Kane、Emily A Scherer 等。2017年。利用移动感知预测精神分裂症症状轨迹。《ACM交互式、移动、可穿戴与泛在技术论文集》 1, 3 2017 1 24 10 1145 3130976 （ ），– 。[122] https://doi.org/ 。/ Rui Wang、Weichen Wang、Mikio Obuchi、Emily Scherer、Rachel Brian、Dror Ben-Zeev、Tanzeem Choudhury、John Kane、Martar Hauser、Megan Walsh 等。2020年。在随机对照试验中利用移动感知预测精神分裂症复发。载于*2020*年*IEEE*普适计算与通信国际会议（*PerCom*）。IEEE，1–8。https://doi.org/10.1109/PerCom45495. 2020.9127365[123] Sun-Chong Wang。2003年。人工神经网络。载于*Java*编程中的跨学科计算。Springer，81–100。[124] Weichen Wang、Subigya Nepal、Jeremy F Huckins、Lessley Hernandez、Vlado Vojdanovski、Dante Mack、Jane Plomp、Arvind Pillai、Mikio Obuchi、Alex daSilva 等。2022年。First-Gen Lens：使用移动感知评估第一代大学生大学第一学年的心理健康。《ACM交互式、移动、可穿戴与泛在技术论文集》6，2（2022），1–32。https://doi.org/ 10 1145 3543194 。 / Xinxi Wang、David [125] Rosenblum 和 Ye Wang。2012。面向日常活动的情境感知移动音乐推荐。载于第*20*届*ACM*国际多媒体会议论文集（日本奈良）（*MM '12*）。计算机协会，美国纽约州纽约市，99–108。https://doi.org/10.1145/ 2393347.2393368[126] Joshua M Wiener、Raymond J Hanley、Robert Clark 和 Joan F Van Nostrand。1990年。日常生活活动的测量：全国调查之间的比较。《老年学杂志》45，6（1990），S229–S237。[127] Garrett Wilson、Janardhan Rao Doppa 和 Diane J Cook。2022年。针对自然时间序列移动活动识别中行为和时间变化的域适应。《arXiv预印本 arXiv:2207.04367》（2022）。[128] Xuhai Xu、Xin Liu、Han Zhang、Weichen Wang、Subigya Nepal、Yasaman Se-fidgar、Woosuk Seo、Kevin S Kuehn、Jeremy F Huckins、Margaret E Morris 等。2023年。GLOBEM：纵向人类行为建模的跨数据集泛化。《ACM交互式、移动、可穿戴与

泛在技术6，4（2023），1–34。[129] Xuhai Xu、Jennifer Mankoff 和 Anind K Dey。2021年。通过被动式移动感知，了解人类状态建模研究人员的实践与需求。《CCF普适计算与交互汇刊》3，4（2021），344–366。[130] Rong Yang 和 Baowei Wang。2016。PACP：一种使用智能手机传感器的位置无关活动识别方法。《Information》7，4（2016年12月），72 10 3390 7040072 。[131] https://doi.org/ 。/info Nira Yuval-Davis。。性别与民族。Routledge。2004。[132] Jingwen Zhang、Dingwen Li、Ruixuan Dai、Heidy Cos、Gregory A Williams、Lacey Raper、Chet W Hammill 和 Chenyang Lu。2022年。利用可穿戴设备预测术后并发症：胰腺手术患者案例研究。《ACM交互式、移动、可穿戴与泛在技术论文集》6，2（2022），1–27。[133] Shichao Zhang。2012。用于迭代式k近邻插补的最近邻选择。《系统与软件杂志》85，11（2012），2541–2552。[134] Wanyi Zhang、Qiang Shen、Stefano Teso、Bruno Lepri、Andrea Passerini、Ivano Bison 和 Fausto Giunchiglia。2021年。将人类行为可预测性置于情境。*EPJ* 数据科学10，1（2021），42。[135] Chen Zheng、Wendy Yajun Huang、Sinead Sheridan、Cindy Hui-Ping Sit、Xiang- Ke Chen 和 Stephen Heung-Sang Wong。2020年。新冠肺炎疫情使年轻成年人过上久坐的生活方式：一项横断面与纵向研究。国际环境研究与公共卫生杂志17，17（8月 2020 6035 10 3390 17176035 ），。https://doi.org/。/ijerph[136] Yuchao Zhou、Suparna De、Wei Wang、Ruili Wang 和 Klaus Moessner。2018年。移动感知环境中的缺失数据估计。*IEEE Access* 6（2018），69869–69882。[137] Ran Zhu、Zhuoling Xiao、Ying Li、Mingkun Yang、Yawen Tan、Liang Zhou、Shuisheng Lin 和 Hongkai Wen。2019 年。通过深度集成学习高效识别人类活动并解决易混淆活动。*IEEE Access* 7（ 2019 75490 75499 10 1109 2019 2922104 ），–。[138] https://doi.org/。/ACCESS.。Zhendong Zhuang 和 Yang Xue。。使用智能手表检测和识别运动相关的人体活动。《传感器》19，22（2019年11月），5001。https: // doi.org/10.3390/s19225001